JIRKM | **Journal of Information Retrieval and Knowledge Management**
Volume 2, 2012

The Journal Secretariat would like to express their heartfelt appreciation to all contributors and reviewers involved in the publishing process of this journal.

e-ISSN: 2289-2664

Published by

Society of Information Retrieval and Knowledge Management

Content

SUPPORTING CONTEXT-AWARE RECOMMENDATIONS: SYSTEM MODEL AND AN EFFICIENT SEQUENTIAL PATTERN MINING ALGORITHM

Jiahong Wang[1]and Eiichiro Kodama[2] and Toyoo Takada[3]
Faculty of Software and Information Science, Iwate Prefectural University, Japan
wjh@iwate-pu.ac.jp

and
Jie Li[4]
Graduate School of Systems and Information Engineering, University of Tsukuba, Japan

Abstract. This paper addresses the subject of context-aware recommendations, and focuses on the mining method of context-related sequential patterns for supporting context-aware recommendations. A typical recommendation system answers such questions as what are the interesting items for the current users. Most traditional recommendation systems have not taken the situational information into account when making recommendations, which seriously limits their effectiveness in ubiquitous computing application environment, where a user's request is generally related to, and thus processing a request should be dependent on, specific contexts (e.g., a specific location, time slot, noise level, or temperature range). This paper proposes a context-aware recommendation system model to improve performance of recommendation systems, which is characterized by a novel sequential pattern mining algorithm that can efficiently mine and group patterns by contexts. Extensive experiments have been conducted, and experiment results demonstrated the effectiveness of the proposed approach.

Keywords: Recommendation system, sequential pattern mining, context-aware computing.

1. Introduction

A recommendation system is a system that provides recommendations, predictions, or opinions to a user on the basis of the user's observed behavior or the behavior of other users. It can be helpful in e-commerce for recommending products, in corporate intranets for providing assistance in finding expertise, or in medical applications where, e.g., patients are matched to doctors.

The context is becoming an important issue for recommendation systems. Traditionally, context is referred to the situation that is related to the user who is accessing the recommendation system, and a system makes recommendations by taking the context information into consideration. For one example, for better services, Amazon should provide different recommendations for the high school student user and graduate school student user who are searching for, for example, CG books. For another example, customer reviews provide customer-to-customer recommendations in Amazon. Users were found to reflect the reviewers' context (e.g., expertise) against their own, and a negative review was typically not seen as a deterrent if the reviewer's context were different from the user's [17].

e-ISSN: 2289-2264

In this paper, we address a different kind of contexts: the context of the activities related to a user's query or a query's results. In other words, we consider the semantic of a user's query. An example is given in Example 1 for explanation. Another example can be found in [21], where the user's experience of an item takes place where the recommendation system itself is running, i.e. on the user's mobile device, and context data can be obtained from sensors of the device.

Example 1. *Refer to Fig. 1. Let us consider an automobile production line, which is divided into several segments called locations. Each location is equipped with several sensors for monitoring its status. Sensors are connected with each other via wireless network and to a monitoring center.*

When an abnormality occurs with a location, maintainers can access the corresponding sensors to learn what is out of order and how it can be dealt with. Such practices are logged routinely. By analyzing the log, experiences in maintaining the production line are accumulated, and a decision tree as shown in Fig. 1 can be constructed. In turn, a recommendation system can be constructed for supporting the maintaining activity.

Figure 1 tells that, each maintaining activity occurs in a specific location, i.e., a location is the context of a maintaining activity. Then novices may ask help to the recommendation system by specifying a context, and recommendation system makes its recommendations in that context.

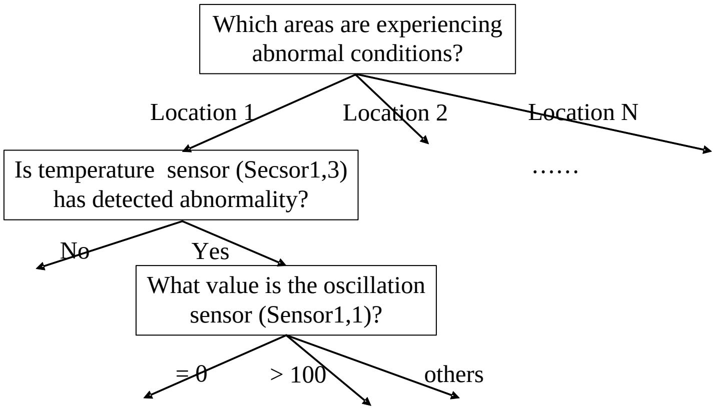

Which areas are experiencing abnormal conditions?
Location 1
Location 2
Location N
Is temperature sensor (Secsor1,3) has detected abnormality?
……
No
Yes
What value is the oscillation sensor (Sensor1,1)?
= 0
> 100
others

Fig. 1: A decision tree for illustrating the concept of per-context sequential patterns.

As described in the above example, in this paper we address the subject of making context-aware recommendations. We achieve the goal by proposing a per-context sequential pattern mining algorithm called *ContextSPM*. ContextSPM can discover sequential patterns related to each context in one pass. Existing sequential pattern mining algorithms such as GSP [1], PrefixSpan [2, 3], SPAM [4, 5], Spirit [6], and SPADE [7] mine a full set of the patterns satisfying a support threshold. Algorithms in [8, 9, 10] mine a full set of the patterns containing no subpatterns with the same support. Algorithms in [3, 4] can mine the patterns beginning with a user-specified item. To the best of our knowledge, there have not been efficient alternatives to the proposed algorithm ContextSPM.

On the basis of ContextSPM we propose a framework for making context-aware recommendation. A user may submit to recommendation system a query with explicitly specifying a context identifier, or the recommendation system may detect context of a user's query. According to the context information, recommendation system makes as appropriate as possible recommendations to a user. The framework is also characterized by a context-aware prefetch cache for enhancing the efficiency of data accesses.

A similar work can be found in [11], where a new sequential access pattern mining algorithm is proposed, on which a web recommendation system is discussed. The context issue, however, is not taken into consideration. Another can be found in [18], where a context-aware recommendation system is proposed that has a context learning service for enabling construction of users' contexts from their log files. Mining the accompanying sequential patterns, however, is not addressed. Similarly, in [20] identifying context information embedded in consumer reviews is studied, and a technique for detecting review's sentences containing the contextual information is proposed. For others, authors of [12] argued that, traditional recommendation systems cannot provide the best recommendation due to lacking the awareness of contexts, and although many research projects have utilized the potentials of context-awareness, the composition of context is fixed at design time, as that in [13]. Thus they proposed a framework that enhances the recommendation systems with a task-oriented, context-aware model. Authors of [22] proposed a recommendation system for recommending the advertisements of the location-based services with consideration of the mobile phone users' location, time, and needs type. Authors of [19] proposed a context-aware layered approach to recommendation, which can abstract raw context information to a semantic level to apply to recommendation system. Different from these work, we emphasize the context-related sequential pattern mining algorithm. We think that it is the mining algorithm that determines the behaviour and the performance of a recommendation system.

The remainder of this paper is structured as follows. In Section 2, we describe the model of computation and formally define the problem. An algorithm for per-context sequential pattern mining, the ContextSPM, is given in Section 3. We also use an example to illustrate ContextSPM. A framework for supporting context-aware recommendation is given in Section 4. Results of a performance study are discussed in Section 5. Finally, Section 6 concludes the paper.

2. Problem Statement

Let $I = \{e_1, e_2, \ldots, e_m\}$ be a set of all items, $C = \{c_1, c_2, \ldots, c_r\}$ be a set of context identifiers, and $C \subseteq I$. A sequence, denoted by $<e_1'e_2'\ldots e_n'>$, is an ordered list of items, where $e_i' \in I$, and $e_i' \in C$ may hold. The length of a sequence is defined as the number of instances of individual items in the sequence. A sequence of length *l* is called a *length-l sequence*.

A sequence $\alpha = a_1 a_2 \ldots a_{l_\alpha}$ is called a subsequence of another sequence $\beta = b_1 b_2 \ldots b_{l_\beta}$, denoted as $\alpha \subseteq \beta$, if there exist integers $1 \leq j_1 < j_2 < \cdots < j_{l_\alpha} \leq l_\beta$, such that $a_i = b_{j_i}$ for all $i \leq l_\alpha$.

A sequence database, denoted by *SD*, is a set of tuples <*sid*, *S*>, where *S* is a sequence, *sid* is its identifier. The support count of a sequence α in *SD* is the number of tuples containing α in *SD*, i.e.,

$$support_{SD}(\alpha) = |\{< sid, S > \ | < sid, S > \in SD \ \wedge \ \alpha \subseteq S\}|$$

The support of a sequence α in *SD* is defined as

$$sup_{SD}(\alpha) = \frac{support_{SD}(\alpha)}{|\ SD\ |}$$

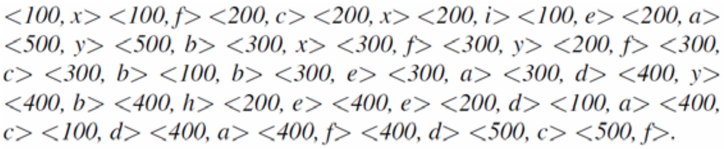
Given a real number *MinSup* ($0 \leq MinSup \leq 1$) as the support threshold, a sequence α is called a *frequent sequential pattern* in *SD* if $sup_{SD}(\alpha) \geq MinSup$. In this case, α is called a *MinSup-pattern*, or a *frequent pattern*, or simply, a *pattern*.

Problem Statement. *The problem of making context-aware recommendations is addressed in two steps.*

a) Given an access log database as the sequence database SD, a list of context identifiers, and a support threshold MinSup. Devise an algorithm to mine all the MinSup-patterns from SD, clustered by context identifiers.

b) Devise a framework of a recommendation system to make the context-aware recommendation on the basis of the mined sequential patterns in Step a).

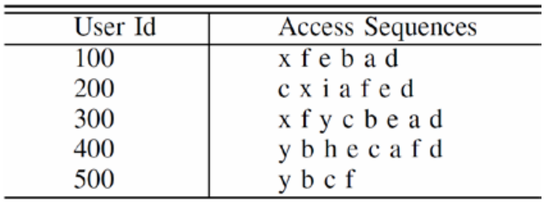
Example 2. *Continuing with Example 1, let us consider the following access log database. Each record in the log is of the format <UserId, Item>. Items "x" and "y" are the context identifiers.*



<100, x> <100, f> <200, c> <200, x> <200, i> <100, e> <200, a> <500, y> <500, b> <300, x> <300, f> <300, y> <200, f> <300, c> <300, b> <100, b> <300, e> <300, a> <300, d> <400, y> <400, b> <400, h> <200, e> <400, e> <200, d> <100, a> <400, c> <100, d> <400, a> <400, f> <400, d> <500, c> <500, f>.

We can divide the log data into access sequences according to UserId. The resulting access sequence database is shown in Table 1. There are totally 5 access sequences. Subsequence xfe is a 60%-pattern because it gets supports from 3 access sequences of user 100, 200, and 300.

Table 1: A Sequence Database SD



User Id	Access Sequences
100	x f e b a d
200	c x i a f e d
300	x f y c b e a d
400	y b h e c a f d
500	y b c f

There are two context identifiers: "x" and "y". We therefore can divide the context-related patterns into two sets: the ones containing "x", and the ones containing "y". Then users' requests concerning "x" and "y" can be answered in the corresponding context.

3. Per-Context Sequential Pattern Mining

This section first introduces the related definitions and preliminaries by using Examples 1 and 2 (Let MinSup = 0.6). Then the algorithm ContextSPM is presented.

3.1 Definitions and Preliminaries

Given the sequence database *SD* shown in Table 1. There exist two kinds of context, identified by items "*x*" and "*y*", respectively. Sequential patterns per context can be mined as follows.

Support counts are first calculated: {*x*:3, *y*:3, *a*:4, *b*:4, *c*:4, *d*:4, *e*:4, *f*:5, *h*:1, *i*:1}. Erasing the infrequent items, we have list *FIList* = <*x, y, a, b, c, d, e, f*>, and *CXList* = <*x, y*>. Then, we can divide the set of patterns in *SD* into the following 3 disjoint subsets:

a) the ones that contain *x*,

b) the ones that contain *y* but do not contain *x*, and

c) the ones that contain neither *x* nor *y*.

Accordingly, the context-related patterns can be mined by constructing two new databases derived from *SD*:

a) the sequences containing *x*, and

b) the sequences containing *y*, but item *x* has been removed.

Motivated by this idea, the problem of mining frequent patterns can be decomposed into a set of subproblems as follows:

Lemma 1 (Level 1 problem partitioning (By the context ids)). *Given a sequence database SD. Let FIList =* $\{x_1, x_2, \ldots, x_n, v_{n+1}, v_{n+2}, \ldots, v_m\}$ *be a complete list of the frequent items, and CXList =* $\{x_1, x_2, \ldots, x_n\}$ *be a complete list of the frequent context identifiers. The set of context-related patterns in SD can be divided into n disjoint subsets. The ith subset, denoted by PSetT(*x_i*), includes the patterns that contain* x_i*, but do not contain* x_j $(j < i)$.

According to this lemma, the complete set of context-related patterns in *SD* can be mined by identifying *PSetT(*x_i*)* $(i \leq n)$ separately in order. For example, in Examples 1 and 2, the patterns can be mined by identifying *PSetT(x)* = {*x, xa, xd, xe, xf, xad, xed, xfd, xfe, xfed*} and *PSetT(y)* = {*y, yb, yc*}, respectively.

Definition 1 (Target item, Seed, and Domain). *Assume we are identifying PSetT(*x_i*).* x_i *is called the target item. The complete set of such sequences in SD that contain* x_i*, but* x_j $(j < i)$ *has been removed, is called* x_i*'s domain, denoted by Domain(*x_i*). The set* $\{x_i, x_{i+1}, \ldots, x_n, v_{n+1}, v_{n+2}, \ldots, v_m\}$ *is called a seed set with regard to* x_i*, denoted by Seed(*x_i*).*

For Examples 1 and 2, we have two domains: *Domain(x)* = {*xfebad, cxiafed, xfycbead*} and *Domain(y)* = {*fycbead, ybhecafd, ybcf*}.

Given a target item, say item *x*, *PSetT(x)* can be mined in the *Domain(x)*. At that time, we need to know the complete set of such items that head patterns in *PSetT(x)*, which is called the *FIRST* item set with regard to item *x*.

Definition 2 (FIRST item set). *Given a target item x. Let PSetT(x) =* $\{\alpha_1, \alpha_2, \ldots, \alpha_m\}$*. Item set* $\{f_1, f_2, \ldots, f_n\}$ *is called the FIRST item set with regard to x, denoted by FIRST(x), if and only if (1) for any* $i \leq n$*, there exists* $j \leq m$*,* $f_i = \alpha_j[1{:}1]$*; (2) for any* $i \leq m$*, there exists* $j \leq n$*,* $\alpha_i[1{:}1] = f_j$*. The notation such as* $\alpha_j[1{:}1]$ *is defined below.*

Given a sequence α= e_1e_2 … e_{l_α}, and e_k is an item in α. A sequence β = $e_1'e_2'$ … e_k' is called a prefix of α with regard to e_k, denoted by $\alpha[1:k]$ or $\alpha[e_k]$ for simplicity, if and only if e_i' = e_i for $i \leq k$. In the case of $e_j \neq e_k$ ($k < j \leq l_\alpha$, or $k = l_\alpha$), $\alpha[e_k]$ is called the maximal prefix, denoted by $\alpha[e_k]_{max}$.

For example, given the item *c* and a sequence α = *abcdecfg*, both *abc* and *abcdec* are its prefixes, represented by α*[1:3]* and α*[1:6]*, respectively. The latter is the maximal prefix, denoted by $\alpha[c]_{max}$.

For Examples 1 and 2, we have *FIRST(x)* = {*x*}, and *FIRST(y)* = {*y*}. Using *FIRST* we can further conduct the second level problem partitioning.

Lemma 2 (Level 2 problem partitioning (By the *FIRST* items)). *Let x be a target item, and FIRST(x) = {f_1, f_2, …, f_m}. PSetT(x) can be divided into m disjoint subsets. The ith subset contains the patterns with prefix f_i, denoted by PSetF(f_i).*

For Examples 1 and 2, take item *x* as the target item. Since *FIRST(x)* = {*x*}, according to this lemma, *PSetT(x)* can be divided into only one subset: *PSetF(x)* = {*x, xa, xd, xe, xf, xad, xed, xfd, xfe, xfed*}.

PSetF is generated by generating length-(l+1) patterns from a length-l pattern that has been mined so far. The items that qualify for attaching to the length-l pattern, called the *qualified items*, are the items that are supported sufficiently in the projected database of length-l pattern.

Definition 3 (Projected database). *Let α be a pattern in Domain(x). The α-projected database, denoted by $Domain(x) \mid_\alpha$, is the set of suffixes of sequences in Domain(x) with regard to α. The term suffix is defined as follows.*

Given a sequence φ = e_1e_2 … e_n. Let β = a_1a_2 … a_m be a subsequence of φ, and a_m = $e_{m'}$ ($m \leq m'$). Sequence γ = $e_{m'+1}e_{m'+2}$ … e_n is called the suffix of φ with regard to β, denoted by $\gamma = \varphi / \beta$.

For example, given φ = *abcdecfg* and β = *abc*, *decfg* is the suffix of φ with regard to β, represented by *abcdecfg* / *abc*. For Examples 1 and 2, take *x* as the target item, and *f* as the α, then *Domain(x)* $\mid_f$ = {*ebad, ed, ycbead*}.

To make sure each pattern in *PSetF(x)* includes the target item, and also to be efficient, the concepts of *candidate set* and *search space* are introduced in Lemma 3.

Lemma 3 (Candidate set and Search space). *Let x_i be the target item, and α a length-l pattern in Domain(x_i). α^+ = {β_1, β_2, …, β_m}, the complete set of length-(l+1) patterns with prefix α, is to be generated. The qualified items, QIS(α) = {β_1/α, β_2/α, …, β_m/α}, need to be determined. The set of subsequences where QIS(α) is searched is called the search space, denoted by SS(α). The set* *of items that can be members of QIS(α) is called the candidate set, denoted by CS(α). Candidate set and search space are determined in four cases (called "states").*

a) *Find-FIRST (α = Null): The FIRST(x_i) is to be determined. SS(Null) = $\{\gamma_1[x_i]_{max}, \gamma_2[x_i]_{max}, \ldots, \gamma_p[x_i]_{max}\}$, where $\{\gamma_1, \gamma_2, \ldots, \gamma_p\}$ = Domain(x_i). CS(Null) = Seed(x_i). FIRST(x_i) is equal to the complete set of frequent items in SS(Null), which is a subset of CS(Null).*

b) *Pre-Target ($x_i \nsubseteq \alpha$): SS(α) = SS(Null) $|_{\alpha}$. CS(α) = QIS(α^-), where α^- denotes the length-(l-1) prefix of α.*

c) *Is-Target ($x_i \nsubseteq \alpha[1:l\text{-}1] \wedge \alpha[l:l] = x_i$): SS($\alpha$) = Domain($x_i$) $|_{\alpha}$, and CS(α) = Seed(x_i).*

d) *Post-Target ($x_i \subseteq \alpha[1:l\text{-}1]$): SS($\alpha$) = Domain($x_i$) $|_{\alpha}$. CS(α) = QIS(α^-).*

For Examples 1 and 2, take item *x* as the target item, *Domain(x)* = {*xfebad*, *cxiafed*, *xfycbead*}. When generating *FIRST(x)*, we have *SS(Null)* = {*x*, *cx*} and *CS(Null)* = {*x*, *y*, *a*, *b*, *c*, *d*, *e*, *f*}. Searching *SS(Null)* we find *FIRST(x)* = {*x*}. Since *x* is the target item, we enter the State 3, and have *SS(x)* = {*febad*, *iafed*, *fycbead*} and *CS(x)* = *CS(Null)*. Searching *SS(x)* we obtain length-2 patterns {*xa*, *xd*, *xe*, *xf*}. Here we have extended the short pattern *x* to longer ones *xa*, *xd*, *xe*, and *xf*, which is justified as follows.

Lemma 4 (Generate PSetF). *Let f be a member of a given FIRST item set. A pattern with prefix f belongs to PSetF(f) if and only if it is generated by the following two steps and it is in either Is-Target or Post-Target state. (1) f is the length-1 pattern. (2) Let α be a length-l pattern ($l \geq 1$), and QIS(α) = $\{q_1, q_2, \ldots, q_p\}$. Then $\{\alpha q_1, \alpha q_2, \ldots, \alpha q_p\}$ is the complete set of length-(l+1) patterns with the prefix α.*

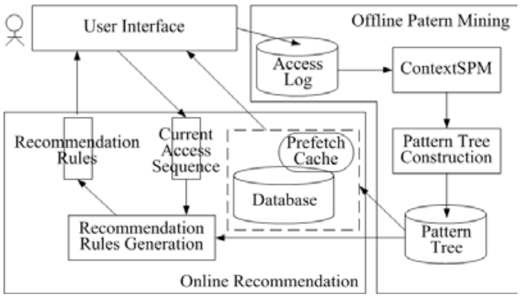
3.2 The Algorithm ContextSPM

Algorithm ContextSPM is given in Fig. 2. Its correctness follows from the above definitions and lemmas.

4. A Framework for Supporting Context-Aware Recommendation

A ContextSPM-based framework for context-aware recommendation is shown in Fig. 3.

As the basis of making recommendations, beforehand a Pattern Tree as shown in Fig. 4 is constructed off-line in three steps. Firstly, users' access activities are logged in an Access Log database. Secondly, ContextSPM is applied to mine access patterns. Lastly, from the mined access patterns the Pattern Tree Construction component constructs a Pattern Tree, and stores it in the Pattern Tree database. Then the system is ready for making recommendations.



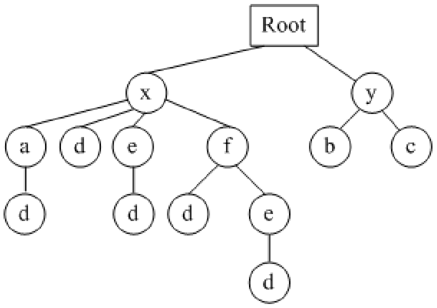
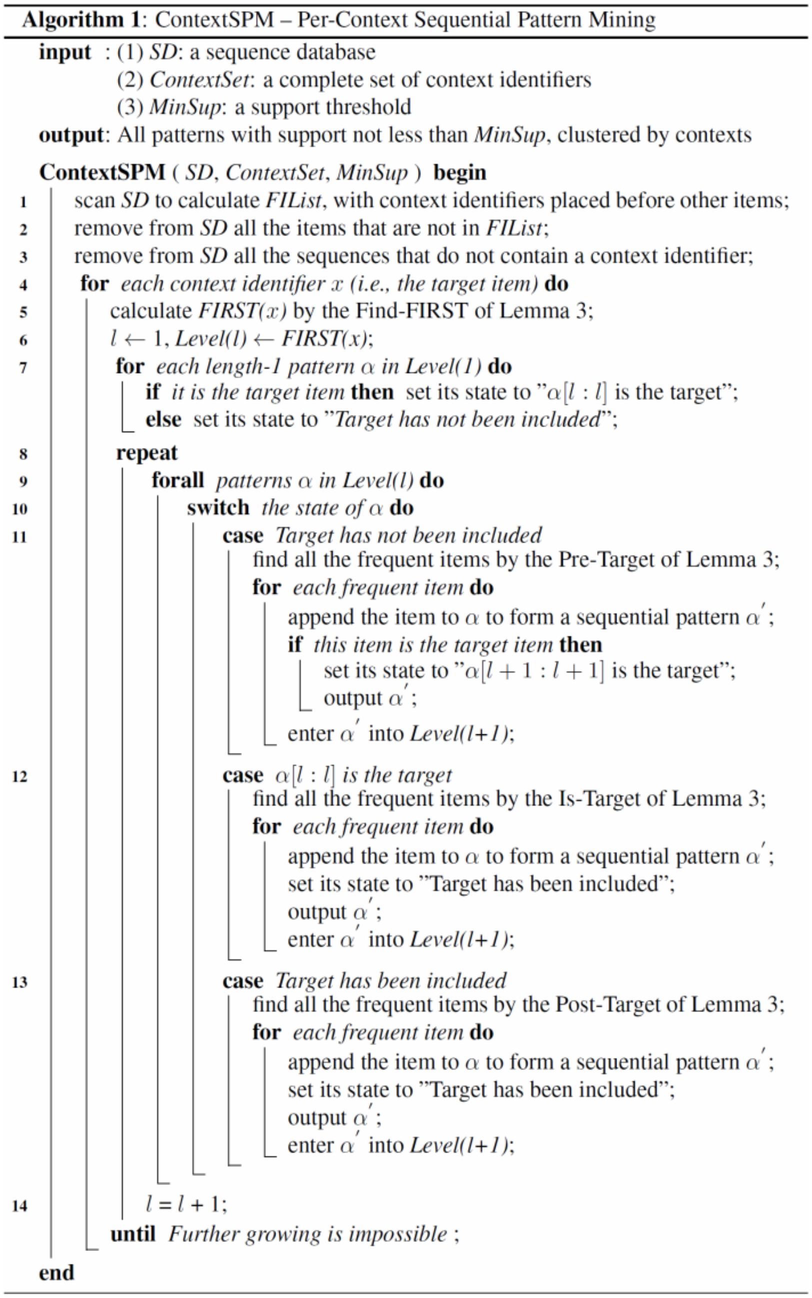


Fig. 3: A framework for supporting context-aware **Fig. 4: Pattern tree for**
Examples 1& 2.
recommendation.

Taken as the Current Access Sequence of a user who is accessing a system, his/her requests in the current session are recorded in order. Matching Current Access Sequence with Pattern Tree, the Recommendation Rule Generation component will generate Recommendation Rules.

The Prefetch Cache component constructs the context-aware Prefetch Cache from Pattern Tree in two steps.

a) A prefetch plan is made for each context by scanning the context's subtree from right to left and from bottom to up, recording each item but ignoring all the reoccurrences. For example, for context "*x*" in Fig. 4, the prefetch plan is *xafed*. By doing so, the order of items in the subtree can be preserved as far as possible, and thus, the next data of a user's access sequence



```
Algorithm 1: ContextSPM – Per-Context Sequential Pattern Mining
input  : (1) SD: a sequence database
         (2) ContextSet: a complete set of context identifiers
         (3) MinSup: a support threshold
output: All patterns with support not less than MinSup, clustered by contexts

ContextSPM ( SD, ContextSet, MinSup ) begin
    scan SD to calculate FIList, with context identifiers placed before other items;
    remove from SD all the items that are not in FIList;
    remove from SD all the sequences that do not contain a context identifier;
    for each context identifier $x$ (i.e., the target item) do
        calculate FIRST($x$) by the Find-FIRST of Lemma 3;
        $l \leftarrow 1$, Level($l$) $\leftarrow$ FIRST($x$);
        for each length-1 pattern $\alpha$ in Level(1) do
            if it is the target item then set its state to "$\alpha[l : l]$ is the target";
            else set its state to "Target has not been included";
        repeat
            forall patterns $\alpha$ in Level($l$) do
                switch the state of $\alpha$ do
                    case Target has not been included
                        find all the frequent items by the Pre-Target of Lemma 3;
                        for each frequent item do
                            append the item to $\alpha$ to form a sequential pattern $\alpha'$;
                            if this item is the target item then
                                set its state to "$\alpha[l + 1 : l + 1]$ is the target";
                                output $\alpha'$;
                            enter $\alpha'$ into Level($l$+1);
                    case $\alpha[l : l]$ is the target
                        find all the frequent items by the Is-Target of Lemma 3;
                        for each frequent item do
                            append the item to $\alpha$ to form a sequential pattern $\alpha'$;
                            set its state to "Target has been included";
                            output $\alpha'$;
                            enter $\alpha'$ into Level($l$+1);
                    case Target has been included
                        find all the frequent items by the Post-Target of Lemma 3;
                        for each frequent item do
                            append the item to $\alpha$ to form a sequential pattern $\alpha'$;
                            set its state to "Target has been included";
                            output $\alpha'$;
                            enter $\alpha'$ into Level($l$+1);
            $l = l + 1$;
        until Further growing is impossible ;
end
```

Fig. 2: Algorithm ContextSPM

can be fetched more efficiently. A prefetch plan can be a by-product of Pattern Tree constructing.

b) The cache is loaded according to prefetch plans. Then, the prefetch cache is put into operation.

Like existing prefetch caches, the efficiency of the proposed prefetch cache depends on whether it matches an application's use of data. We found that in the domain of sensor networks such as what is given in Examples 1 and 2, access sequences tend to be repeatable, and the proposed prefetch cache can function very efficiently.

To construct Pattern Tree from the mined access patterns, in Pattern Tree Construction component an n-way tree [14] is utilized.

The Recommendation Rule Generation component searches for the best matching access path in Pattern Tree according to a user's Current Access Sequence, and provide the suffix for a user as the recommendation.

5. Performance Evaluation

Considering the difference between the proposed framework and its alternatives is mainly in the sequential pattern mining component, we will evaluate ContextSPM only.

Because we have not found a complete alterative to ContextSPM, we cannot directly compare it with others. We take PrefixSpan [3] as a reference, since it is an influential and one of the fastest sequential pattern mining algorithms. For the comparison, we have modified ContextSPM to take every item as a target item, and therefore to mine all the patterns, as PrefixSpan does.

Note that it has been reported by other previous studies that, PrefixSpan is faster than SPAM [15] and other earlier algorithms such as those found in [16, 1, 7].

5.1. Experiment Environment

Synthetic sequence databases, generated using the IBM data set generator from the IBM Almaden Research center, were used.

Performance evaluation was conducted in terms of

a) *support threshold*,

b) *sequence database size*, measured in number of sequences, denoted by letter *D*,

c) *average length of sequences*, denoted by letter *S*,

d) *average length of potential patterns*, denoted by letter *I*, and

e) *number of individual items, denoted by letter N.*

All experiments were conducted on a 1.5GHz Pentium PC with 1GB memory.

5.2. Experiment Results and Discussions

The *execution time*, defined as the time interval from the mining process is started to it is finished, is used as a performance metric.

Figure 5 shows execution times with support threshold varying from 0.05% to 0.25%. ContextSPM is found to outperform PrefixSpan. The maximal memory usages of PrefixSpan and ContextSPM are 6.7MB and 234MB, respectively, implying that compared with the PrefixSpan, ContextSPM can (1) utilize memory resources more reasonably and effectively, and (2) may be limited if one has not sufficient memory resources. We do not think this more usage of memory resources would limit ContextSPM, since unlike before, modern computers generally have a large amount of memory resources available.

Figure 6 shows execution times with database size growing from 100K to 500K sequences. ContextSPM still outperforms PrefixSpan. But PrefixSpan outperforms ContextSPM slightly in the scalability in this case, since execution time with ContextSPM increases slightly faster. The difference in scalability was found to be small, and in the case of a heavy environment (e.g., longer average lengths or lower support thresholds), they had almost the same scalability.

Figure 7 shows execution times with average length of the sequences growing from 5 to 25 items. For both algorithms, the longer the average length is, the larger the execution time is. However, comparatively ContextSPM is far less affected.

For examining the behaviours of ContextSPM under different environments, an experiment with different databases has been done with varying support thresholds from 0.25% to 0.05%. The results are shown in Fig. 8. This figure tells that for a heavy environment, a too small threshold would give very long execution times. Therefore, if the access logs are very large, or if one has not enough computing resources such as for the mobile phones, it is not encouraged to use a small threshold setting. In fact, large threshold settings always mean more helpful recommendations.

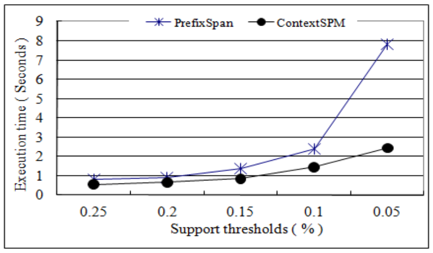


Fig. 5: Support thresholds vs. Execution times Execution Times (Database is S10I4D100KN10000). S10I4

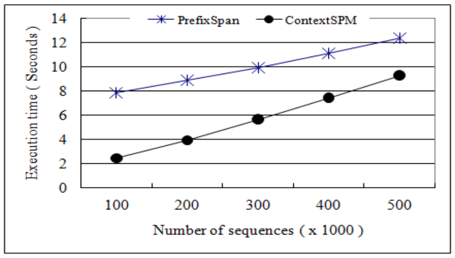


Fig. 6: Number of sequences vs.

(Support is 0.05%, and database is

D100-500KN10000)

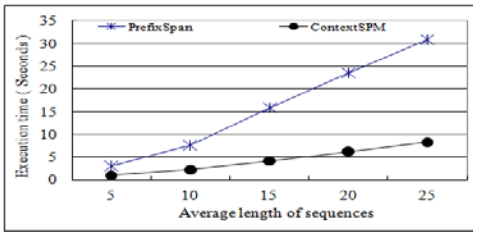


Fig. 7: Average Length of Sequences vs. times Execution Times (Support is 0.05%. Database is S5-25I4D100KN10000).

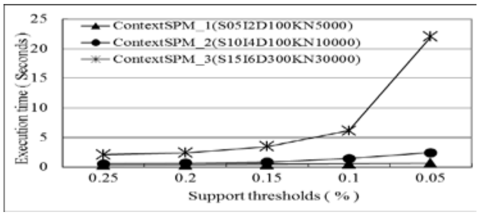


Fig. 8: Support thresholds vs. Execution

6. Conclusions

In this paper we have addressed the subject of making context-aware recommendations. A per-context sequential pattern mining algorithm, called ContextSPM, has been proposed. On the basis of ContextSPM, a framework for context-aware recommendation systems has also been proposed. In the framework, ContextSPM is used to analyze the historical information to identify relations between items and then these relations are used to determine the recommendations. ContextSPM is characterized by its ability to mine sequential patterns in every context in one pass. The results are represented in a pattern tree, with each context's patterns being represented within one subtree.

Extensive experiments have been conducted to compare ContextSPM with PrefixSpan. The experimental results demonstrated that ContextSPM outperforms PrefixSpan in most cases. It was also found that, compared with PrefixSpan, ContextSPM requires more memory space. Therefore, for the applications where the memory is extremely limited, ContextSPM would not be a proper candidate. This requirement, however, would not be so difficult to meet since the memory of modern computers is becoming cheaper and larger.

References

[1] R. Srikant and R. Agrawal. 1996. Mining sequential patterns: Generalizations and performance improvements, In *Proc. the 5th International Conf. on Extending Database Technology (EDBT'96)*, pp. 3-17, Avignon, France.

[2] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, and M.C. Hsu. 2001. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth, In *Proc. the 17th International Conf. on Data Engineering*, pp.215-224, Heidelberg, Germany.

[3] J. Pei, J. Han, B. Mortazavi-Asl, J. Wang, H. Pinto, Q. Chen, U. Dayal, and M.C. Hsu. 2004. Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach, *IEEE Transactions on Knowledge and Data Engineering*, Vol.16, No.11, pp.1424-1440.

[4] J. Ayres, J. Gehrke, T. Yiu, and J. Flannick. 2002. Sequential Pattern Mining Using Bitmaps, In *Proc. the 8th ACM SIGKDD International Conf. on Knowledge Discovery and Data Mining*, pp.429-435, Edmonton, Alberta, Canada.

[5] Z. Yang and M. Kitsuregawa. 2005. LAPIN-SPAM: An Improved Algorithm for Mining Sequential Pattern, In *Proc. 2005 International Special Workshop on Databases For Next Generation Researchers (SWOD'05) in conjunction with ICDE'05*, pp. 8-11, Tokyo, Japan.

[6] M. N., Garofalakis, R. Rastogi, and K. Shim. 1999. Spirit: Sequential pattern mining with regular expression constraints, In *Proc. the 25th International Conference on Very Large Data Bases*, pp. 223-234, Edinburgh, Scotland, UK.

[7] M.J., Zaki. SPADE: An efficient algorithm for mining frequent sequences, *Machine Learning*, Vol.42, No.1/2, pp.31-60.

[8] X. Yan, J. Han, and R. Afshar. 2003. CloSpan: Mining Closed Sequential Patterns in Large Datasets, In *Proc. the 3rd SIAM International Conference on Data Mining (SDM03)*, pp. 166-177, San Francisco, USA.

[9] J. Wang and J. Han. 2004. BIDE: Efficient Mining of Frequent Closed Sequence, In *Proc. the 20th International Conf. on Data Engineering (ICDE)*, pp. 79-90, Boston, USA.

[10] Z. Li, Z. Chen, S. Srinivasan, and Y. Zhou. 2004. C-Miner: Mining Block Correlations in Storage Systems, In *Proc. the 3rd USENIX Conference on File and Storage Technology (FAST'04)*, pp. 173-186, San Francisco.

[11] B. Zhou, S. Hui, and A. Fong. 2006. Efficient Sequential Access Pattern Mining for Web Recommendations, *International Journal of Knowledge-Based and Intelligent Engineering Systems*, Vol.10, No.2, pp. 155-168.

[12] G. Yap, A. Tan, and H. Pang. 2005. Dynamically-Optimized Context in Recommender Systems, In *Proc. the 6th International Conf. on Mobile Data Management*, pp.265-272, Ayia Napa, Cyprus.

[13] G. J. F. Jones and P. J. Brown. 2003. Context-aware Retrieval for Ubiquitous Computing Environments, In *Proc. 2003 Mobile and Ubiquitous Information Access Workshop*, pp. 227-243, Udine, Italy.

[14] K. Peeters. *An STL-like C++ N-Way Tree Class*, from http://www.aei.mpg.de/peekas/tree/.

[15] K. Wang, Y. Xu, and J. Yu. 2004. Scalable Sequential Pattern Mining for Biological Sequences, In *Proc. the 13th ACM international conf. on Information and knowledge management (CIKM'04)*, pp. 178-187, Washington, D.C., USA.

[16] R. Agrawal and R. Srikant. 1995. Mining sequential patterns, In *Proc. the 11th International Conf. on Data Engineering*, pp. 3-14, Taipei, Taiwan.

[17] J. Leino and K.J. Räihä. 2007. Case Amazon: Ratings and Reviews as Part of Recommendations, In *Proc. the 2007 ACM conference on Recommender systems (RecSys '07)*, pp. 137-140, Minneapolis, Minnesota, USA.

[18] S. Abbar, M. Bouzeghoub, and S. Lopes. 2009. Context-Aware Recommender Systems: A Service Oriented Approach, In *Proc. the 3rd International Workshop on Personalized Access, Profile Management and Context Awareness in Databases (PersDB'09), in conjunction with the VLDB Conference*, Lyon, France.

[19] D. Shin, J.W. Lee, J.H. Yeon, and S.G. Lee. 2009. Context-Aware Recommendation by Aggregating User Context, In *Proc. 2009 IEEE Conference on Commerce and Enterprise Computing (CEC'09)*, pp. 423-430, Vienna, Austria.

[20] S. Aciar. 2010. Mining Context Information from Consumer's Reviews, In *Proc. 2nd Workshop on Context-Aware Recommender Systems (CARS'10), Barcelona, Spain.*

[21] M. Bohmer, G. Bauer, and A. Kruger. 2010. Exploring the Design Space of Context-aware Recommender Systems that Suggest Mobile Applications, In *Proc. 2nd Workshop on Context-Aware Recommender Systems (CARS'10), Barcelona, Spain.*

[22] K.J. Kim, H. Ahn, and S. Jeong. 2010. Context-aware Recommender Systems using Data Mining Techniques, In *Proc. World Academy of Science, Engineering and Technology (WASET'10)*, Rome, Italy.

LECTURE TIMETABLING USING IMMUNE-BASED ALGORITHMS

Muhammad Rozi Malim
Faculty of Computer and Mathematical Sciences, University Technology MARA, 40450 Shah Alam, Malaysia
rozi@tmsk.uitm.edu.my

Abstract: Lecture timetabling is a highly constrained optimization problem. Metaheuristic approaches, and their hybrids, have successfully been applied to solve the problem. This paper presents three immune-based algorithms for lecture timetabling; clonal selection, immune network, and negative selection. The ultimate goal is to show that the algorithms may be adapted as new alternatives for solving lecture timetabling problems. The algorithms have been implemented on four benchmark lecture (class) datasets. Experimental results have shown that all algorithms are good optimization algorithms; have successfully produced good quality lecture timetables. The algorithms are compared based on fitness values, relative robustness, and CPU times. Statistical tests of hypotheses have significantly shown that the immune network is more effective than the other two algorithms. All algorithms can handle the hard and soft constraints very well, and may be accepted as new members of evolutionary algorithms for timetabling. The values of relative robustness have shown that the timetables produced by clonal selection are more robust than those produced by other two algorithms. The recorded CPU times have revealed that the immune network has acquired the longest time on all datasets. A comparison with published results has shown that all algorithms are as good as other solution methods. For future work, these algorithms will be employed to other domains of timetabling problems.

Keywords: *Lecture Timetabling; Artificial Immune System; Immune-Based Algorithm.*

1. Introduction

Various facets of biology have always been the inspiration in developing computational models and problem solving methods. The rapid increase in research of the biological systems has enabled us to gain insight into the miraculous operation of our body. The use of biologically inspired metaphors can result in new computer technologies and methods of problem solving, and computing can provide new techniques for exploring biological concepts from an alternative prospective. The *immune system* (IS), a biological system, has recently drawn significant attention; and as a result, the *artificial immune system* (AIS) has emerged. In 1986 the theoretical immunologist, J. D. Farmer, first suggested a possible relationship between immunology and computing [12]. Since then, the field has expanded rapidly, with numerous papers published applying AIS to a diverse set of topics ranging from computer security [13] to robotics [20].

The potential application areas of the IS metaphors are those seeking robust and good-enough solutions to problems occurred in dynamic environments [17]. These features are characteristic of a number of real-world problem domains; anomaly detection, pattern recognition, computer/network security, dynamic environments, dynamic learning, robotics, diagnosis and control, data analysis, optimization, and scheduling. The task of producing *robust schedules* (large similarity) has a direct analogy with the task faced by the IS; both operate in a dynamic and unpredictable environment. The AISs are relatively new techniques and have been successfully applied to optimization and scheduling problems [2, 3, 10, 18, 19, and 24]. Since timetabling is a special case of *scheduling*, and treated as *optimization* by the Operations

Research community, the AIS approaches may be adapted for timetabling problems to produce good quality (and robust) timetables.

Lecture timetabling problem (LTP) is a specific case of the more general timetabling problem, and known to be a highly constrained optimization problem. The LTP, also called *course* or *class* timetabling, can be viewed as a multi-dimensional assignment problem. Given a set of courses, a set of teachers, a set of weekly timeslots, a set of classrooms, and a set of student enrollments, the problem is to assign teachers to courses, and lectures of courses to timeslots and classrooms satisfying a set of hard and soft constraints. Many different approaches, including evolutionary algorithms (EAs), Tabu search (TS), simulated annealing (SA), and their hybrids are developed for solving many different types of LTPs.

Hard constraints must be satisfied to produce a feasible timetable, whilst violation of soft constraints should be minimized and provides a measure of how good the solution is via an objective (*fitness*) function. Soft constraints are generally more numerous and varied and far more dependent on the needs of the individual problem than the more obvious hard constraints. The LTP can be seen as consisting of three subproblems; *course-teacher*, *lecture-timeslot*, and *lecture-room* assignments. In course-teacher assignment, the teachers are scheduled to all lectures of courses; in lecture-timeslot assignment, all lectures are scheduled into a limited number of timeslots; and in lecture-room assignment, all lectures are assigned to a fixed number of rooms. Hence, in a LTP, an assignment is an ordered 4-tuple (a, b, c, d), where $a \in E$, $b \in T$, $c \in R$, and $d \in P$. An assignment has the straightforward general interpretation: event (lecture) a starts at timeslot b in room c, and is taught by teacher d. For some institutions, the allocation of courses to teachers is carried out manually, and the allocation of lectures to rooms is a secondary problem and can be done later as a separate activity.

This paper presents *three* immune-based algorithms (or AIS algorithms) for lecture timetabling; *clonal selection*, *immune network*, and *negative selection*. The ultimate goal is to show that the three algorithms may be adapted as new alternative approaches for solving LTPs. *Four* benchmark lecture timetabling datasets are used for the implementation. Another objective is to compare the effectiveness of the three algorithms on lecture (class) datasets. The algorithms are compared based on the fitness values (soft constraint violations), the CPU times, and the relative *robustness* (similarity between timetables). A comparison with published results is also conducted to show that the algorithms are comparable with other solution methods.

2. Problem Statement and Objectives

Timetabling problems are known to be highly complex scheduling problems. These problems are always studied because of its variety and complexity. The increase in the number of students (and courses) and complexity of program structures mean that timetables are becoming more complex and difficult to schedule. New timetables must be produced for every single semester to take account of staff, student and course changes, causing a large amount of administrative work. Small changes in timetabling data would ruin the feasibility of a timetable. AIS approaches may be applied to produce *robust timetables* (large similarity timetables). With a set of robust timetables, another feasible timetable (suite the changes) may be selected. In scheduling and AIS, the term *robustness* is synonym with the flow-shop and job-shop scheduling problems [19 and 22] but not timetabling.

Heuristic methods are often used to solve real-world timetabling problems. The most popular and well-studied heuristics are *metaheuristics* which include simulated annealing (SA), Tabu search (TS), evolutionary algorithms (EAs). However, many of these approaches *lack the robustness*. AIS schedules are *robust* and observed to be robust than schedules produced by a standard *genetic algorithm* [19]. However, no timetabling researchers have applied AIS algorithms. AIS and Timetabling are two separate disciplines, and there already exist a large number of good timetabling algorithms.

The aim of this paper is to introduce a number of immune-based algorithms for lecture timetabling. The ultimate goal is *not* to show that the algorithms are better than other well-established algorithms in timetabling, but to show that the algorithms may be adapted as new alternative approaches for solving timetabling problems.

The main objectives are as follows:

(a) Propose *immune-based algorithms* for lecture timetabling problems based on *three immunological principles* (clonal selection, immune network, and negative selection); i.e. *three* AIS algorithms are considered.

(b) Formulate lecture timetabling problem as a mathematical model.

(c) Implement and compare the AIS algorithms on benchmark timetabling datasets.

(d) Compare the results with other timetabling approaches.

3. Methodology

This section contains information on the approaches and methods employed to realize the objectives on which this paper is based. The first section (3.1) outlines a model for lecture timetabling, the timetabling constraints, a mathematical approach for problem formulation, and the quality of a feasible timetable. The second section (3.2) looks at the adaptation of AIS algorithms for lecture timetabling, mapping between AIS and lecture timetabling, and the main operators of AIS algorithms. Section 3.3 summarizes the implementation procedure of the algorithms on lecture timetabling datasets, a measure for robustness, the techniques used to compare the algorithms including the statistical tests of hypotheses, and the comparison with published results.

3.1 Lecture Timetabling Model

Nine lecture timetabling variables are considered; Department (D), Academic-program (M), Course (C), Student-group (G), Student (S), Staff (P), Event (lecture) (E), Timeslot (T), and Room (R). The interrelationships connecting the nine variables are represented by matrices. The *matrix representation* is widely used in timetabling to organize a large and complex data. There are two types of matrices, *input* and *output*. The number of input matrices depends on the available timetabling data while the number of output matrices depends on the desired timetables. An input matrix shows the association between two variables, and an output matrix shows a timetable.

Based on the nine variables, the hard and soft lecture timetabling constraints are formulated using *0-1 integer programming* (0-1 IP) approach. 0-1 IP is the special case of integer programming where variables are required to be 0 or 1 (rather than arbitrary integers). In contrast to linear programming, which can be solved efficiently in the worst case, 0-1 IP problems are in many practical situations *NP-hard* (nondeterministic polynomial-time hard), i.e. no method of solving it in a reasonable (polynomial) amount of time is known. Also, the formulation uses a logic function $L(*, *, *)$; a function of timetabling variables, always has a logic value of either 0 or 1. This function is suitable for formulating complex timetabling constraints. Each logic function is unique and can only be applied for a particular constraint.

For timetabling problems with a large number of hard constraints, or *over-constrained* problems, obtaining timetables that satisfy all the hard constraints at once is reasonably difficult. To overcome this, a *constraint relaxation* mechanism is usually employed. The relaxed hard constraints are considered as soft constraints and will be satisfied during an *improvement process*. One must introduce a sort of hierarchy upon relaxed constraints. A *weight* is then proposed for each constraint. This weight allows the setting of a partial order relation between constraints. The lower the value of the weight is the more important the constraint is.

The *quality* of a feasible timetable is usually determined by a set of different soft constraints. The constraints are formed as a *fitness function*; it represents the total violations of the soft constraints (to be minimized). A *penalty* (weight) *function* is required for each of the soft constraints. This function uniquely reflects the importance of each soft constraint in the objective function. The most important soft constraint (the most desirable to be satisfied) would acquire the highest penalty value. A timetable with *lowest fitness* value is always considered as the *best quality* timetable. However, different institutions have different needs and requirements, and hence have different ways of determining a quality timetable.

3.2 Adapting AIS Algorithms for Lecture Timetabling

The *immune system* (IS) can be considered to be a remarkably efficient and powerful information processing system which operates in a highly parallel and distributed manner [17]. It contains a number of features which can be adapted in computer systems; recognition, feature extraction, diversity, learning, memory, distributed detection, self-regulation, threshold mechanism, co-stimulation, dynamic protection, and probabilistic detection. From the perspective of information processing, it is unnecessary to replicate all of these aspects in a computer model, rather they should be used as general guidelines in designing a system.

This section presents a procedure for developing immune-based algorithms for lecture timetabling. *Three* different immunological principles are considered (clonal selection, immune network, and negative selection), and hence three different algorithms are developed. It is important at this stage to establish an appropriate mapping between the IS and LTP (Table 1). For some terms, different immunological principles use different names.

Table 1: Mapping between 'IS' and 'LTP'

Immune System	Lecture Timetabling (LTP)
Antibody/immune cell/detector	Feasible timetable
Gene	Event (lecture)
Antigen	Cloned & mutated feasible timetable
Antigen detectors / lymphocytes	Initial population of feasible timetables
Monoclonal antibodies	Duplicate/identical timetables
Clonal deletion	A process to remove duplicate timetables
Clone (offspring antibody)	Cloned timetable
Cloning	An operator to produce cloned timetables
Mutation	An operator to transform cloned timetables into new feasible timetables
Genetic variation	A process to produce new better quality timetables using cloning & mutation
Receptor editing	Mutation on a number of lectures
Affinity / Stimulation level	Inverse of a fitness value
Network	Population
Metadynamics (antigens and genetic variations)	A process to produce a population of mutated feasible timetables
Network dynamics (immune cells and antigens interactions)	A process to produce a new population of better quality feasible timetables

A feasible timetable (*antibody*, *immune cell*, or *detector*) is an integer-valued vector, i.e. a string of lectures (*genes*). An initial population is a set of different feasible timetables (*antigen detectors* or *lymphocytes*). A duplicate timetable (*self-reactive immature lymphocyte*) is removed by cloning and mutation (*clonal deletion*). The *fitness* value (to be minimized) determines how good a feasible timetable is. The *affinity* (to be maximized) is inversely proportional to the fitness value. Good timetables (low fitness) are selected for cloning and mutation (*genetic variation*) using affinity. The primary objective of the *cloning* is to emphasize good timetables and eliminate bad timetables. So, overall fitness of a population becomes better. Cloning operator copies good timetables from the current population to the next generation population. *Mutation* transforms identical timetables (*monoclonal antibodies*) into different new feasible timetables.

As claimed by previous researchers, algorithms based upon the clonal selection principle are adequate to solve optimization and scheduling problems [8, 14, and 18]. However, the other two principles were also applied to solve the same type of problems [1 and 6]. There still not enough evidence to conclude that clonal selection is the best principle for optimization and scheduling. Timetabling may be considered as another case of optimization and scheduling problems. The AIS is new in timetabling and no one can claim which immune principle is the best for timetabling. Since there are three different principles, three different immune-based algorithms for lecture timetabling can be developed. The clonal selection algorithm (CSA) is inspired by the clonal selection principle [7], the immune network algorithm (INA) is based on the immune network theory [21], and the negative selection algorithm (NSA) is developed using the negative selection mechanism [7].

(a) Initialization Phase of the Immune-based Algorithms

All three immune-based algorithms for LTPs use the *same* initialization phase. In this phase, an initial population of feasible timetables (satisfy all hard constraints) is generated using *a heuristic* or *a set of heuristics* depending on the difficulty of the hard constraints. Some of the heuristics are *largest degree*, *saturation degree*, *largest weighted degree*, *largest enrolment*, *largest number of papers*, *user defined priority groups*, and *random ordering*. These heuristics order the events in some way and attempt to allocate each event (lecture) to a timeslot (or a number of timeslots), satisfying all the hard constraints.

For each feasible timetable, a course is selected (one by one) until all lectures of all courses have been scheduled. All lectures for each course are assigned to staff, timeslots and rooms without violating the hard constraints. If the timetable is *unique* (no duplicates), it is added to the initial population; otherwise, it is eliminated. The process is repeated until the number of feasible timetables in the initial population is equal to the population size. As in the natural IS, the initial population is a large number of *lymphocytes*, each with a different specificity of *antibody*. A duplicate timetable is a potential *self-reactive immature lymphocyte*.

(b) Improvement Phase, Stopping Criteria, & Operators

The improvement phase is an *iterative* optimization process. *Optimization process* is the discipline of adjusting a process so as to optimize some specified set of parameters (objective function) without violating some constraints. The most common goals are minimizing cost, maximizing throughput, and/or efficiency. At each generation (iteration), the current population of feasible timetables is improved (optimized) to produce a better (quality) population using the *inspired immune principle*. The *quality* of each timetable is measured via a *fitness* (objective) function. This process is repeated until some *stopping criteria* are met. Generally the stopping criteria are the *maximum number of generations* and the *maximum number of none improvement generations*.

The main *reproduction operators* in the improvement phase for all immune-based algorithms for lecture timetabling are *cloning* and *mutation*. Cloning operator copies good timetables from the current population to the next generation population. It is expected that the timetables with lowest fitness (greatest affinity) will have more clones. Hence, the number of clones for a timetable is proportional to its *affinity*. Mutation transforms duplicate (cloned) timetables into a number of different feasible timetables with better fitness. The algorithm may be converging upon a local optimum, and mutation is a way to avoid getting stuck in local optimum.

The *multipoint mutation* operator is considered for all immune-based algorithms. It works by randomly selecting a small number of lectures, e.g. 1% or 2% of the total number of lectures, and *reassign* the lectures to the best available timeslots and rooms (minimize fitness), satisfying all the hard constraints. If the number of selected lectures is larger, say 5% or more, the CPU time that acquired to complete the mutation would increase exponentially. A small percentage of mutated lectures would maintain the large similarities between timetables (*robust population*). As in the natural IS, only a small number of genes of an antibody are mutated (*receptor editing*) such that maintaining the *robustness* of antibodies. Each cloned timetable is mutated according to a *mutation probability*. The mutation probability of a cloned timetable is inversely proportional to its affinity. The higher the affinity is the lower the mutation probability.

3.3 Implementation Procedure of Immune-based Algorithms

This section summarizes how the three immune-based algorithms are implemented, tested, and compared on benchmark lecture timetabling datasets. The algorithms (with LTPs formulated as 0-1 IP models) are encoded into *C++ programming language*. Each algorithm is an optimization timetabling algorithm. The results (the feasibility of timetables and the fitness values) are *validated* at the end of each algorithm.

(a) Benchmark Lecture Timetabling Datasets

Four *Schaerf* lecture (class) timetabling datasets are used to implement, test, and compare the three immune-based algorithms. The problems are defined and formulated as 0-1 IP models. These datasets provide a number of benchmark problems for comparison of various timetabling algorithms, and available on the internet from *www.diegm.uniud.it/ satt/projects/EduTT/CourseTT* (*original version*).

(b) Population Size, Number of Generations, and Number of Trials

Choosing the *population size* is a fundamental decision faced by all optimization algorithms researchers, and varies from 10 to 1000. If too small a population size is selected, the algorithms will converge too quickly, with insufficient processing of too few schemata [15]. On the other hand, a population with too many results in long waiting times for significant improvement. The relatively *small populations* are appropriate for *serial implementations* and *large populations* are appropriate for perfectly *parallel implementations*. Koljonen and Alander [23] confirmed the common belief that decreasing population size increases optimization speed to a certain point, after which premature convergence slows the optimization speed down. The optimization reliability in turn usually increases monotonically with increasing population size. For timetabling algorithms, with complex and variety hard constraints, a population size of 10 feasible timetables is more appropriate. For some problems, it is difficult and time consuming even to get five feasible timetables. All implementations of the three algorithms in Section 5 use population size 10, and this remains constant in all generations.

The *number of generations* (or iterations) seemed to matter less than the population size. Selecting the number of generations for which an algorithm is run is often *trial-and-error* process [11]. In general, given enough computing time, the number of generations is adjusted until the desired response is obtained. Other factors, such as population diversity and fitness improvement of the best population member, can enter into the decision to end the run. For example, if the best fitness has not changed for, say, 100 generations, we may choose to terminate the run (*idle iterations*). The optimum number of generations is often a function of the problem. However, for the three immune-based algorithms on Schaerf datasets, the maximum number of generations is fixed to 1000 for all trials. This number would give enough time for the algorithms to converge and optimize the fitness. The fixed number of generations would enable us to compare the CPU times of the algorithms. A greater number of generations, say 2000 generations, would produce better fitness for some problems, but the processing time would increase exponentially.

The *number of trials* for each immune-based algorithm on each dataset is depending on the number of different initial populations that the algorithm can produce. For some datasets, it is difficult to produce even one initial population. For this reason, only 10 trials are produced by each algorithm on each dataset. Hence, each algorithm is required to generate 10 initial populations for each dataset. However, only *five* trials (with best fitness) are selected as the final results.

(c) Measuring the Robustness

Hart et al. [19] defined *robustness* as similarity of two schedules and introduced a measure that based on *Hamming Distance*. The similarity of two schedules is directly compared by counting the number of events that the two schedules differ. For the three immune-based algorithms, the robustness is measured for a population of feasible timetables, and not for a single timetable. For each final population produced from each algorithm, the robustness of the population is measured as follows:

(i) Compare the similarity of each feasible timetable with other timetables of the same population. For population size 10, there are 45 (${}^{10}C_2$) comparisons.

(ii) For each comparison (two timetables), count the number of lectures with different timeslots and rooms.

(iii) Calculate the total differences by summing the values obtained in (b) from all comparisons. This gives a measure of robustness for a final population.

However, the value calculated in (iii) is a large number for a large timetabling problem. To compare the robustness of two populations of two different timetabling problems, a *relative measure* of robustness must be determined. This measure should take into account the number of lectures and the population size. The following measure of robustness for a population of feasible timetables is seemed more reasonable:

$$\textit{Relative Robustness} = \frac{\sum_{i=1}^{n-1}\left(\sum_{j=1}^{m} L(e_{i,j}, e_{i+1,j})/m\right)}{{}^{n}C_2}, \qquad (1)$$

where $L(e_{i,j}, e_{i+1,j})$ = 1 if event (lecture) e_j is assigned at different timeslots/rooms in timetables i and i+1, 0 otherwise; m is the number of lectures, and n is the population size.

Equation (1) actually represents the *average relative differences* of feasible timetables in a population. Hence, a population with the *lowest* relative robustness (largest similarity) is considered as the *most robust*. Now the robustness of populations of timetables produced by different algorithms may be compared.

(d) Comparing Immune-based Algorithms

The immune-based algorithms are compared based on the fitness values, the relative robustness, and the CPU times. A fitness value represents the total violations of the soft constraints. For each dataset, an algorithm with the *lowest* fitness is considered as the *most effective* algorithm. The average fitness is also calculated and compared for each dataset based on the best five trials. The values of relative robustness for all immune-based algorithms on each dataset are calculated using equation (1) and compared to determine which algorithm has produced the most robust population of timetables. Note that, the objective of solving timetabling problems is to produce good quality timetables (low fitness), and *not* to optimize the robustness. The comparison of CPU times, the time elapsed for 1000 generations, would indicate which algorithm is the fastest (or slowest). The CPU times (in seconds) are recorded on an Intel Celeron-M Processor 370 1.5GHz 1.24GB RAM Notebook PC.

(e) Statistical Tests of Hypotheses

The averages of the fitness values of the three immune-based algorithms are further compared using the statistical test of hypotheses. Assuming the populations of the fitness values of the three algorithms on all datasets are normally and independently distributed, the two-tailed *t*-tests (small samples) may be applied to compare the averages of the fitness values. Three *t*-tests were carried out to compare three averages (three algorithms) for each dataset. The sample size is n = 5 (trials). For each test, the *null hypothesis* (H_0) is 'two averages are equal' and the *alternative hypothesis* (H_1) is 'two averages are not equal'.

Since the population variances are unknown, the test statistic t and the degrees of freedom υ are given by

$$t = \sqrt{n}(\overline{X}_1 - \overline{X}_2) / \sqrt{s_1^2 + s_2^2} \ , \tag{2}$$

$$\upsilon = (n-1)(s_1^2 + s_2^2)^2 / [(s_1^2)^2 + (s_2^2)^2] \quad ; \quad \text{(rounded to the nearest integer)} \tag{3}$$

where $\overline{X}_1, s_1^2$ and $\overline{X}_2, s_2^2$ are the sample means and variances of the fitness values of two algorithms being compared, and n is the sample size.

Each t-statistic (2) is compared with $t_{\alpha/2,\nu}$ (from t-distribution table) with α = 0.05 (5% significance level). If t-statistic is less than the corresponding $t_{\alpha/2,\nu}$, then the null hypothesis is *not* rejected, and it may be concluded that the two averages are equal; otherwise, the null hypothesis is rejected and the two average are not equal. These statistical tests of hypotheses are carried out on each dataset.

(f) Comparison with Published Results

The best fitness and the average fitness values for all algorithms and datasets are compared with *published results* [9] to show that the immune-based algorithms are comparable with other solution methods. The results are compared based on the fitness values, and there are no results available on robustness and CPU times.

4. Immune-Based Algorithms

This section presents three immune-based algorithms for lecture timetabling; clonal selection, immune network, and negative selection. These algorithms are developed based on the immunological principles and the AIS algorithms developed by previous researchers, and the methodology described in Section 3. Each algorithm may be applied to solve a wide range of lecture timetabling problems.

Each of the three algorithms mainly consists of two phases; *initialization* and *improvement*. The initialization phase has been discussed in Section 3.2(a). In the improvement phase (generation loop), an AIS approach is performed. This is an iterative process; for each generation, while some stopping criteria are not met, the current population of feasible timetables is improved to produce a better (quality) population. The main improvement operators are *cloning* and *mutation*; cloning produces exact copies of timetables (clones), while mutation transforms the clones into new (improved) feasible timetables, and hence both ensure the success of each generation. The quality of each timetable is measured via a fitness function.

4.1 Clonal Selection Algorithm for Lecture Timetabling

Figure 1 illustrates the *Clonal Selection Algorithm for Lecture Timetabling* (CSALT). This algorithm is developed based on the clonal selection principle [7], an AIS algorithm proposed by Doyen et al. [10], and CLONALG developed by de Castro [4]. As described in Section 3.2(a), an initial population of feasible timetables is generated using heuristics. Then the improvement phase is performed. This is an iterative optimization process. At each generation (iteration), the current population is improved (optimized) to produce a better (quality) population of feasible timetables using the clonal selection principle. The quality of each timetable is measured via a fitness function. This process is repeated until some stopping criteria are met.

(a) Affinity Evaluation

The optimization process, to improve the quality of the current population of feasible timetables, starts with affinity evaluation. An affinity function plays an important role in AIS because it is used to decide how good a solution is. In natural IS, the affinity value determines how good an antibody recognizes foreign antigens. The affinity is inversely proportional to the fitness value. The fitness function can clearly be made up of any timetabling related factors (soft constraints). For each feasible timetable of the current population, a fitness value is determined via a fitness function (the total violations of soft constraints). Then the affinity of each timetable, the total affinity of all timetables, and the minimum affinity (and its associated timetable) are determined for the current population.

```
1. Initialization Phase:  [as described in Section 3.2(a)]
2. Improvement Phase:  While stopping criteria are not met
                          For each timetable of the current population
   Affinity evaluation:     Determine the affinity of each timetable (affinity = 1/fitness)
                          Determine the total affinity of all timetables and the minimum affinity
   Selection:             Calculate the selection probability for each timetable
                            (= affinity/total affinity)
                          Construct the cumulative selection probabilities
                          Generate a random probability
                            Select a timetable based on cumulative probabilities
   Genetic variation:     Clone copies of the selected antibody
                          (no. of clones=population size x cumulative selection probability)
   (cloning & mutation)   Determine a mutation probability (= 1 - selection probability)
                          For each generated clone (cloned timetable)
                            Generate a random probability
                            If a random probability ≤ mutation probability (failure)
                              While mutation = failure
                                Mutate - multipoint mutation (select a small no. of lectures and
                                reassign, satisfying all hard constraints)
                                If no duplicate timetables
                                  determine the affinity of new clone
                                  If the affinity ≥ affinity (original)
                                     mutation = success
                                  else mutation=failure, reset mutation
                                else mutation=failure, reset mutation
   Population update:         If affinity (new clone)> min affinity, say X, replace X with new clone
                              else eliminate the new clone
                            else no mutation & eliminate current clone
```

Fig. 1: Clonal Selection Algorithm for Lecture Timetabling (CSALT)

(b) Selection Operator

A feasible timetable is randomly selected to be the basis of the next generation. Good timetables (high affinity) are more likely to be chosen than bad ones. By weighting the selection process in favor of the better quality timetables, the worse timetables are eliminated. The selection method is based on selection probabilities. The *selection probability* for each timetable is calculated as follows:

$$\textit{Selection probability} = (\textit{affinity of timetable})/(\textit{total affinity}). \qquad (4)$$

The *cumulative* selection probabilities for all timetables are then constructed. A random probability is generated and then one timetable is selected for cloning based on the cumulative probabilities.

(c) Genetic Variation (Cloning and Mutation Operators)

Good timetables are selected for genetic variation (cloning and mutation) using affinity. The primary objective of the cloning is to emphasize good timetables and eliminate bad timetables. So, overall affinity of a population becomes better. Cloning operator copies good timetables from the current population to the next generation population. It is expected that the timetables with greater selection probabilities will have more clones. The number of clones of the selected timetable is determined by:

$$\textit{Number of clones} = \textit{population size} \times \textit{cumulative selection probability}. \quad (5)$$

When an antibody interacts with an antigen, the cell is activated and starts to divide. It gives rise to clones of identical receptor (monoclonal antibodies). Mutation transforms monoclonal antibodies into antibodies each with unique specificity. Hence, the clones generated from the selected timetable need 'mutation' to 'remove' duplicate timetables and improve their affinity. The algorithm may be converging upon a local optimum, and mutation is a way to avoid getting stuck in local optimum. The effect of mutation is to reintroduce divergence into a convergence population. As described in Section 3.2, the (multipoint) mutation operator works by randomly selecting a small number of lectures (1% or 2% of the total number of lectures), and reassign the lectures to the best available timeslots and rooms (minimize fitness), satisfying all the hard constraints. Each cloned timetable is mutated according to a *mutation probability*, calculated as:

$$\textit{Mutation probability} = 1 - \textit{selection probability}. \quad (6)$$

For each cloned timetable, a random probability is generated. A (feasible) mutation is performed if the probability is less than or equal to the mutation probability; otherwise, the clone is eliminated. If there are no duplicate timetables in the current population, the affinity of the mutated timetable is then determined; otherwise, the mutation process is repeated. If the affinity is greater than or equal to the affinity of the original clone, the mutation is successful. Otherwise, the cloned timetable must repeat the mutation until successful.

(d) Population Update

For each successful mutated clone (new clone), if the affinity is greater than the minimum affinity (of a feasible timetable) of the current population, then the new clone will replace the minimum affinity timetable; otherwise, the new clone is eliminated. A new population of feasible timetables for the next generation is produced when all clones have been mutated, with one or more new feasible timetables. This new population will undergo the same improvement process until the stopping criteria are met.

4.2 Immune Network Algorithm for Lecture Timetabling

Figure 2 illustrates the *Immune Network Algorithm for Lecture Timetabling* (INALT). This algorithm is developed based on the immune network theory [21], the general INA by de Castro [4], and the opt-aiNET by de Castro and Timmis [6]. As in CSALT, the algorithm starts by generating an initial population of feasible timetables using heuristics. Then the improvement phase is performed using the immune network theory. While stopping criteria are not met, the initial population is optimized to produce a better (quality) population of feasible timetables. The quality is measured via a fitness function. The stopping criteria are the maximum number of generations and/or the maximum number of none improvement generations.

(a) Network Interactions and Stimulation

A stimulation level plays an important role in INALT; it is used to decide how good a timetable is. In natural immune system, the stimulation level determines how good an immune cell recognizes other cells or foreign antigens. The stimulation level of a feasible timetable is inversely proportional to a fitness function.

$$\textit{Stimulation level} = 1/(\textit{fitness value}). \tag{7}$$

For each feasible timetable of the current population, the fitness is determined via a fitness function, and then the stimulation level is calculated. The stimulation probability for each timetable can now be calculated.

$$\textit{Stimulation probability} = \textit{stimulation level/total stimulation level.} \tag{8}$$

```
1. Initialization Phase:  [as described in Section 3.2(a)]
2. Improvement Phase:  While stopping criteria are not met
                         For each  timetable of the current population
   Network interactions :  Determine the fitness via a fitness function
   and Stimulation level   Calculate the stimulation level (= 1/fitness)
                         Determine total stimulation of current population
                         Calculate stimulation probability for each timetable
                         (= stimulation/total stimulation)
   Metadynamics:         Initialize a population of mutated timetables (antigens)
   (Antigens and         For each feasible timetable
   Genetic variations)     Cloning – generate a number of clones
                           (= population size x stimulation probability)
                           Determine a mutation probability (=1- stimulation probability)
                           For each cloned timetable
                             Generate a random probability
                             If random probability ≤ mutation probability, mutate
                               While mutation = failure
                                 Mutate - multipoint mutation (select a small number of lectures
                                 and reassign, satisfying all the hard constraints)
                                 If no duplicates (original & mutated)
                                   Calculate the stimulation level
                                   If stimulation≥ stimulation (original)
                                      mutation = success
                                     Add timetable to mutated population
                                   else mutation = failure, reset
                                 else mutation = failure, reset
                             else no mutation, eliminate current clone
   Network dynamics:     Gather all feasible timetables (original and mutated)
   (antigens interactions, Sort timetables according to stimulation levels
   & population update)  Select the best (high stimulation) timetables
                         (= population size) to form a new population
                         Replace the original population with the new population
```

Fig. 2: Immune Network Algorithm for Lecture Timetabling (INALT)

(b) Metadynamics (Antigens and Genetic Variations)

A population of mutated feasible timetables (antigens) is now initialized. All timetables (current population) are selected to be the basis of the next generation. Each timetable will reproduce (by cloning and mutation) to generate a better population of (high stimulation) feasible timetables. By weighting the number of clones in favor to the stimulation probabilities, the low stimulation timetables are eliminated. The timetables with greater stimulation probabilities will have more clones. The number of clones and the mutation probability for each feasible timetable are determined by:

$$\textit{Number of clones} = \textit{population size} \times \textit{stimulation probability}, \tag{9}$$

$$\textit{Mutation probability} = 1 - \textit{stimulation probability}. \tag{10}$$

When an immune cell recognizes an antigen or another cell, it stimulates and starts to divide. It gives rise to immune cells of identical receptors. Mutations are required to transforms these cells into cells each with unique specificity. For each cloned timetable, a random probability is generated. A multipoint mutation (as in CSALT) is performed if the random probability is less than or equal to the mutation probability; otherwise, the clone is eliminated. For each mutated timetable, if there are no duplicate timetables in the populations of original and mutated timetables, the stimulation level is then determined; otherwise, the mutation is repeated. If the stimulation level of the mutated timetable is greater than or equal to the stimulation level of the original timetable, the timetable is added to the population of mutated timetables. Otherwise, the mutation remains failure, then the process is repeated. The population of mutated feasible timetables is completed when all feasible timetables of the current population have been cloned and mutated (or eliminated).

(c) Network Dynamics (Immune Cells, Antigens Interactions, and Update)

All feasible timetables of the ‘original’ and ‘mutated’ populations are gathered. Then the timetables are sorted according to their stimulation levels in descending order. The best (high stimulation) timetables (equal to the population size) are selected to form a new population of feasible timetables. Finally, the original population of feasible timetables is replaced with the new population. A new population of feasible timetables for the next generation is now produced, with one or more (or none) new timetables. This new population will undergo the same optimization process (loop) until the stopping criteria are met.

4.3 Negative Selection Algorithm for Lecture Timetabling

Figure 3 illustrates the *Negative Selection Algorithm for Lecture Timetabling* (NSALT). This algorithm is developed based on the negative selection mechanism [7], the standard NSA proposed by de Castro [4], and the RNS algorithm by Gonzalez et al. [16]. As in CSALT and INALT, the algorithm starts by generating an initial population using heuristics. As in the natural IS, the initial population is a large number of antigen detectors (lymphocytes), each with a different specificity of unique antigen receptor. In the improvement phase, while the stopping criteria are not met, the initial population is optimized to produce a better quality population using the negative selection mechanism. The quality is measured via a fitness function.

```
1. Initialization Phase:  [as described in Section 3.2(a)]
2. Improvement Phase:  While stopping criteria are not met
                          For each feasible timetable of the current population
                            Determine the fitness via a fitness function
                          Determine the average fitness of current population
  Censoring:              If all fitness are equal, eliminate the second half of the timetables
                          (half of the population)
                          else For each feasible timetable
                            If the fitness $\geq$ average, eliminate timetable
                            else the timetable remains in current population
  Monitoring:             While number of feasible timetables < population size
                            Randomly select a timetable according to fitness values
                            Clone the timetable (single cloning)
                            While mutation = failure
                              Mutate - multipoint mutation (select a small number of lectures and
                              reassign, satisfying all the hard constraints)
                              If no duplicate timetables, determine the fitness of new timetable
                                If fitness $\leq$ average, mutation=success
                                  Add new timetable to current population
                                else (mutation = failure) reset
                              else (mutation = failure) reset mutation
```

Fig. 3: Negative Selection Algorithm for Lecture Timetabling (NSALT)

(a) Censoring

Censoring process eliminates high fitness feasible timetables according to the average fitness. At each generation, the fitness value is determined for each feasible timetable. Then the average fitness of the current population is calculated as follows:

$$\textit{Average} = \textit{total fitness/population size}. \tag{11}$$

If all fitness values are equal (all timetables would be eliminated), eliminate the second half of the population. Otherwise, each fitness value is compared to the average fitness; if the fitness is greater than or equal to the average, remove the timetable from the population. The remaining timetables must reproduce to generate a new (better fitness) population.

(b) Monitoring

Monitoring process generates new feasible timetables, to replace the eliminated ones, by cloning and mutating the remaining timetables. On average, only half of the feasible timetables remain in the current population. New timetables must be generated so that the number of feasible timetables always equal to the population size. While the number of feasible timetables is less than the population size, randomly select a timetable from the current population. By weighting the selection process in favor of the fitness, it is expected that the low fitness timetables will be selected and reproduced. So the overall fitness of the population becomes better.

For each selected timetable, only ‘one’ clone is produced (*single cloning*). This cloned timetable needs mutation to remove duplicates. The mutation probability is equal to 1, i.e. all selected timetables would be mutated (multipoint mutation). For each mutated timetable, if no duplicates, the fitness is determined; otherwise, the mutation is repeated until successful. If the fitness is less than or equal to the current average, the mutation is a success and the timetable is added to the current population. Otherwise, the mutation is repeated until successful. The monitoring process is repeated until the number of feasible timetables in the current population is equal to the population size. Now a new population of feasible timetables for the next generation is produced (on average, half are new timetables). The improvement phase is repeated until the stopping criteria are met.

5. Implementation and Experimental Results

In this section (5.3 and 5.4), the three immune-based algorithms are implemented and compared on four lecture timetabling benchmark datasets. The characteristics of the datasets are given in Section 5.1. The timetabling problems of the datasets are formulated as 0-1 IP models in Section 5.2.

5.1 Lecture Timetabling Benchmark Datasets and Formulation

The Schaerf lecture (class) timetabling datasets are available at *www.diegm.uniud.it/satt/projects/EduTT/CourseTT* (*original version*). These are the real-world lecture timetabling instances from the School of Engineering at the University of Udine, called *Schaerf datasets*. The datasets and characteristics are shown in Table 2.

Table 2: Schaerf Lecture Datasets and Characteristics

Dataset	No. of Courses	No. of Teachers	Timeslots per day	No. of Rooms (R)	No. of Timeslots (T)	Total timeslots (RXT)	Total lectures (L)	Conflict Density	Occupancy (L/(R×T))
1	46	39	4	12	20	240	207	4.64%	86.25%
2	52	49	4	12	20	240	223	4.75%	92.92%
3	56	51	4	13	20	260	252	4.61%	96.92%
4	55	51	5	10	25	250	250	4.78%	100%

Each dataset comes in *five* files; *courses.dat* contains the information about the courses (course ID, course name, preassigned teacher, number of lectures, minimum number of days, and number of students), *periods.dat* contains the list of two-hour timeslots of the timetabling horizon (day, start time, finish time), *curricula.dat* contains the information about student-groups with courses that share common students (student-group ID, number of courses, and list of courses), *constraints.dat* contains additional constraints about unavailability of timeslots (course ID, day, and unavailable timeslot), and *rooms.dat* contains information about rooms (room ID, and number of seats). The lecture timetabling horizon is of five days; datasets 1, 2 and 3 have four timeslots each day, while dataset 4 has five timeslots each day. The *conflicts density* shows the percentage density of the conflict matrix of courses and curricula, and the *occupancy* shows the percentage of the total timeslot-rooms required to schedule all lectures of all courses. These values indicate the degrees of difficulty of the problems. Obviously, dataset 4 is the most complex and would acquire the longest CPU time.

The Schaerf lecture timetabling problem for all datasets is defined as follows (using the variables as defined earlier). There are n_1 preassigned teachers (staff) p_i, $i \in \{1,...,n_1\}$, n_2 events of courses (lectures) e_j, $j \in \{1,...,n_2\}$, n_3 timeslots t_{k_1}, $k_1 \in \{1,...,n_3\}$, and n_4 rooms r_{k_2}, $k_2 \in \{1,...,n_4\}$. Each lecture e_j has weekly frequency $n_f(e_j)$ to be scheduled in distinct timeslots, and it is attended by $n(e_j)$ students. Each room r_{k_2} has a capacity $n_{cs}(r_{k_2})$, expressed in terms of available seats. There are also n_5 groups of courses (student-groups), called *curricula*, g_{k_3}, $k_3 \in \{1,...,n_5\}$, such that any two lectures of a curriculum have students in common. For datasets 1, 2 and 3, there are 20 timeslots per week of five days (4 slots/day), and dataset 4 has 25 timeslots (5 slots/day).

Five hard constraints are considered; (i) all lectures of all courses must be scheduled, (ii) two distinct lectures cannot take place in the same room and timeslot, (iii) lectures of courses (same curriculum) must be scheduled at different timeslots, (iv) lectures of courses taught by the same teacher must be scheduled at different timeslots, and (v) teachers might be not available for some timeslots. *Three soft constraints* are considered and used to evaluate the fitness value of each feasible timetable; (vi) the number of students that attend a course should be less than or equal to the number of seats of all rooms that host its lectures, (vii) the lectures of each course should be spread into not less than a specified minimum number of days, and (viii) the daily schedule of lectures (same curriculum) should be as compact as possible, avoiding gaps between courses. A 'gap' is a free timeslot between two lectures (same curriculum) scheduled on the same day. The objective is to produce a number lecture timetables such that all the hard constraints are satisfied, and the total violations of the soft constraints are minimized. The penalty of the soft constraint (vi) is the 'number of students without seats', whereas the weights of the soft constraints (vii) and (viii) were fixed to 5 and 2, respectively. The hard constraint violations (if any) are assigned the weight 1000. With this weight, the violated hard constraint(s) would be satisfied in the improvement process.

There are *five* variables (staff, event (lecture), timeslot, room, and student-group). The teachers (staff) have been preassigned to courses (lectures). The problem is to assign lectures of courses to timeslots and rooms, satisfying all the hard constraints and minimizing the total violations of the soft constraints. Using the 0-1 IP approach, the mathematical formulation may be carried out as follows: The input matrices are *student-group–event allocation* (**A**), *staff-event preassignment* (**B**), *event-timeslot restriction* (**C**). The required output matrices (timetables) are *event-timeslot assignment* (**X**) and *event-room assignment* (**Y**). For each dataset, the input matrices are constructed using the given data files, and then the output matrices are constructed by solving the problem.

The 0-1 IP model for the Schaerf lecture timetabling problem, with five hard constraints and three soft constraints, may be formulated as:

minimize

$$\sum_{k_1=1}^{n_3}\sum_{k_2=1}^{n_4} L_{RC}(e_j,t_{k_1},r_{k_2}) + \left(5\times\sum_{j=1}^{n_2} L_{ES}(e_j,t_{k_1})\right)$$

$$+ \left(2\times\sum_{k_3=1}^{n_5}\sum_{D=1}^{5}\sum_{k_1=T_1}^{T_2-2}\left(\sum_{j=1}^{n_2} a_{k_3 j}\cdot x_{jk_1}\right)\cdot\left(1-\sum_{j=1}^{n_2} a_{k_3 j}\cdot x_{j(k_1+1)}\right)\cdot\left(\sum_{j=1}^{n_2} a_{k_3 j}\cdot x_{j(k_1+2)}\right)\right),\tag{12}$$

subject to

$$\sum_{j=1}^{n_2} L_{EF}(e_j,t_{k_1}) = 0\,,\tag{13}$$

$$\sum_{k_1=1}^{n_3}\sum_{k_2=1}^{n_4} L_R(e_j,t_{k_1},r_{k_2}) = 0\,,\tag{14}$$

$$\sum_{k_3=1}^{n_5}\sum_{j_1=1}^{n_2-1}\sum_{j_2=j_1+1}^{n_2} a_{k_3 j_1}\cdot a_{k_3 j_2}\cdot L_{EC3}(e_j,t_{k_1}) = 0\,,\tag{15}$$

$$\sum_{i=1}^{n_1}\sum_{k_1=1}^{n_3} L(e_j,p_i,t_{k_1}) = 0\,,\tag{16}$$

$$\sum_{j=1}^{n_2}\sum_{k_1=1}^{n_3}(1-c_{jk_1})\cdot x_{jk_1} = 0\,,\tag{17}$$

all variables are integers 0-1;

where $L_{RC}(e_j,t_{k_1},r_{k_2}) = [\ \sum_{j=1}^{n_2} n_{\mathbf{G}}(e_j)\cdot y_{(jk_1)k_2}\ \text{-}\ n_{cs}(r_{k_2})\ ]$ if $\sum_{j=1}^{n_2} n_{\mathbf{G}}(e_j)\cdot y_{(jk_1)k_2} > n_{cs}(r_{k_2})$, 0 otherwise; $n_{\mathbf{G}}(e_j)$ is the number of students in lecture event e_j, and $n_{cs}(r_{k_2})$ is the room-capacity for room r_{k_2}; $L_{ES}(e_j,t_{k_1})=1$ if $\sum_{D=1}^{5} L_{ESD}(e_j,t_{k_1}) < n_{\mathrm{mind}}(e_j)$, 0 otherwise; $L_{ESD}(e_j,t_{k_1})=1$ if $\sum_{j=1}^{n_2}\sum_{k_1=T_1}^{T_2} x_{jk_1} \geq 1$, 0 otherwise; $n_{\mathrm{mind}}(e_j)$ is a specified minimum days for lecture event e_j; T_1, T_2 are respectively the first and last timeslots on day D; $L_{EF}(e_j,t_{k_1})=0$ if $\sum_{k_1=1}^{n_3} x_{jk_1} = n_f(e_j)\cdot l(e_j)$, 1 otherwise; $n_f(e_j)$ and $l(e_j)$ are respectively the weekly frequency and the length (hours) of lecture event e_j; $L_R(e_j,t_{k_1},r_{k_2})=0$ if $\sum_{j=1}^{n_2} y_{(jk_1)k_2} \leq 1$, and 1 otherwise; $L_{EC3}(e_j,t_{k_1})=0$ if $\sum_{k_1=1}^{n_3} x_{j_1k_1}\cdot x_{j_2k_1} = 0$, and 1 otherwise; and $L(e_j,p_i,t_{k_1}) = 0$ if $\sum_{j=1}^{n_2} a_{ji}\cdot x_{jk_1} \leq 1$, and 1 otherwise.

5.2 Implementation and Comparison of Immune-based Algorithms on Lecture Timetabling Datasets

In this section, the three immune-based algorithms for lecture timetabling are implemented on four Schaerf lecture datasets. The main objective is to show that each algorithm may produce good quality (low fitness) lecture timetables. Other objectives are to compare the effectiveness of the algorithms on lecture datasets, the relative robustness, and the CPU times. A comparison with a published result [9] is carried out to show that the immune-based algorithms are capable of producing good quality lecture timetables as good as other solution methods.

On all datasets, a *constraint relaxation approach* has been applied since it is almost impossible to generate a population of timetables that satisfy all the five hard constraints at once. For first two datasets (1 and 2), the *third* hard constraint is relaxed (i.e. considered as a soft constraint). This constraint will be satisfied in the improvement phase. Therefore, for each pair of lectures of the same curriculum that scheduled in the same timeslot, a penalty of 1000 is incurred. The new 0-1 IP model for datasets 1 and 2 (with four soft constraints and four hard constraints) may be rewritten as

minimize $$\sum_{k_1=1}^{n_3}\sum_{k_2=1}^{n_4} L_{RC}(e_j,t_{k_1},r_{k_2}) + 5\times\sum_{j=1}^{n_2} L_{ES}(e_j,t_{k_1})$$

$$+$$

$$2\times\sum_{k_3=1}^{n_5}\sum_{D=1}^{5}\sum_{k_1=T_1}^{T_2-2}\left(\sum_{j=1}^{n_2} a_{k_3j}\cdot x_{jk_1}\right)\cdot\left(1-\sum_{j=1}^{n_2} a_{k_3j}\cdot x_{j(k_1+1)}\right)\cdot\left(\sum_{j=1}^{n_2} a_{k_3j}\cdot x_{j(k_1+2)}\right)$$

$$+\ 1000\times\sum_{k_3=1}^{n_5}\sum_{j_1=1}^{n_2-1}\sum_{j_2=j_1+1}^{n_2} a_{k_3j_1}\cdot a_{k_3j_2}\cdot L_{EC3}(e_j,t_{k_1}), \tag{18}$$

subject to $$\sum_{j=1}^{n_2} L_{EF}(e_j,t_{k_1}) = 0, \tag{19}$$

$$\sum_{k_1=1}^{n_3}\sum_{k_2=1}^{n_4} L_R(e_j,t_{k_1},r_{k_2}) = 0, \tag{20}$$

$$\sum_{i=1}^{n_1}\sum_{k_1=1}^{n_3} L(e_j,p_i,t_{k_1})=0\,, \tag{21}$$

$$\sum_{j=1}^{n_2}\sum_{k_1=1}^{n_3}(1-c_{jk_1})\cdot x_{jk_1}=0\,, \tag{22}$$

all variables are integers 0-1.

For the third and fourth datasets, two hard constraints (*second* and *third*) are relaxed. For the second hard constraint, a penalty of 500 is incurred to each room at each timeslot if two or more lectures were scheduled simultaneously. Similarly, for the third hard constraint, a penalty of 500 is added for each pair of lectures of the same curriculum that are scheduled at the same timeslot. Since there are two relaxed hard constraints, instead of 1000, the weight 500 is enough to ensure that the relaxed constraints will be satisfied in the improvement process. The new 0-1 IP model for datasets 3 and 4 (five soft and three hard constraints) may be rewritten as

minimize

$$\begin{aligned}&\sum_{k_1=1}^{n_3}\sum_{k_2=1}^{n_4} L_{RC}(e_j,t_{k_1},r_{k_2}) \;+\; 5\times\sum_{j=1}^{n_2} L_{ES}(e_j,t_{k_1})\\ &+\\ &2\times\sum_{k_3=1}^{n_5}\sum_{D=1}^{5}\sum_{k_1=T_1}^{T_2-2}\left(\sum_{j=1}^{n_2} a_{k_3 j}\cdot x_{jk_1}\right)\cdot\left(1-\sum_{j=1}^{n_2} a_{k_3 j}\cdot x_{j(k_1+1)}\right)\cdot\left(\sum_{j=1}^{n_2} a_{k_3 j}\cdot x_{j(k_1+2)}\right)\\ &+\\ &500\times\left(\sum_{k_3=1}^{n_5}\sum_{j_1=1}^{n_2-1}\sum_{j_2=j_1+1}^{n_2} a_{k_3 j_1}\cdot a_{k_3 j_2}\cdot L_{EC3}(e_j,t_{k_1})+\sum_{k_1=1}^{n_3}\sum_{k_2=1}^{n_4} L_R(e_j,t_{k_1},r_{k_2})\right)\\ &,\end{aligned} \tag{23}$$

subject to

$$\sum_{j=1}^{n_2} L_{EF}(e_j,t_{k_1})=0\,, \tag{24}$$

$$\sum_{i=1}^{n_1}\sum_{k_1=1}^{n_3} L(e_j,p_i,t_{k_1})=0\,, \tag{25}$$

$$\sum_{j=1}^{n_2}\sum_{k_1=1}^{n_3}(1-c_{jk_1})\cdot x_{jk_1}=0\,, \tag{26}$$

all variables are integers 0-1.

For dataset 4, the ‘100% occupancy’ would make the mutation impossible, i.e. not enough rooms to reassign lectures. However, two *dummy rooms* (with zero seats) are introduced to solve the problem. If a lecture is assigned to a dummy room, then all students attended that lecture are considered as *students without seats*. For all algorithms and datasets, the multipoint mutation operator selects and reassigns only 2% of the number of courses. However, this 2% is equivalent to *one* for all datasets since the maximum number of courses is 56; hence, only *one* course is selected. As defined in Section 3.3, the *maximum number of generations* (stopping criterion) for each algorithm and trial is 1000, and the population size is 10. For each dataset, all algorithms are implemented using the same initial populations (one set one trial).

(a) Implementation and Comparison of Immune-based Algorithms on Benchmark Datasets

The CSALT, INALT, and NSALT presented in Sections 4.1, 4.2 and 4.3, respectively, have been implemented on each Schaerf lecture timetabling dataset (10 trials). An initial population of 10 feasible timetables is generated using a *random ordering heuristic*; i.e. a course is randomly selected (one by one) and then all lectures of the course are assigned to random timeslots and rooms, satisfying all the hard constraints. The fitness value of each timetable is evaluated via the fitness function (19) for datasets 1 and 2, or (24) for datasets 3 and 4. This function is equal to *the number of students without seats*, plus *the number of courses that assigned to less than the specified minimum number of days* multiplies by 5, plus *the number of gaps between lectures of the same curriculum on the same day* multiplies by 2, and plus *the total violations of hard constraint(s)* multiplies by 500 or 1000. The multipoint mutation operator selects *one* course at random, and *reassigns* all lectures of the course to random timeslots and the best available rooms (minimize the number of students without seats), always maintaining a feasible timetable. The improvement phase is repeated for 1000 generations. The fitness values, the values of relative robustness, and the CPU times (1000 generations) for the best *five* trials (low fitness) produced by the three immune-based algorithms are summarized in Tables 3, 4, and 5, respectively.

Table 3: Summarized Results on Schaerf Lecture Datasets using CSALT

Dataset	No. of Lectures	No. of Rooms	No. of Timeslots	Best Fitness	Average Fitness	Ave. Relative Robustness	Ave. CPU Time (seconds)
1	207	12	20	244	256.2	0.031798	375.0s
2	223	12	20	25	49.6	0.023976	112.0s
3	252	13	20	35	59.6	0.013898	151.2s
4	250	10+2	25	129	145.8	0.013760	4028.6s

For dataset 1, at 1000th generation, the best fitness is 244; i.e. 200 students without seats, 4 courses scheduled less than the specified minimum days, and 12 free gaps between lectures of the same curriculum on the same day. For datasets 2, 3, and 4, the best fitness values are 25, 35, and 129, respectively. The *minimum* average relative robustness is 0.013760 (dataset 4), i.e. on average 98.62% of the timetables produced by CSALT at 1000th generation on dataset 4 are similar. Obviously, dataset 4 (most conflict) has acquired the longest average CPU time.

Table 4: Summarized Results on Schaerf Lecture Datasets using INALT

Dataset	No. of Lectures	No. of Rooms	No. of Timeslots	Best Fitness	Average Fitness	Ave. Relative Robustness	Ave. CPU Time (seconds)
1	207	12	20	256	264.4	0.148406	529.0s
2	223	12	20	4	7.8	0.153383	204.2s
3	252	13	20	27	37.2	0.159630	242.0s
4	250	10+2	25	108	119.6	0.034098	9185.0s

For dataset 1, the best fitness is 256; i.e. 200 students without seats, 6 courses scheduled less than the specified minimum days, and 13 free gaps between lectures of the same curriculum on the same day. For datasets 2, 3, and 4, the best fitness values are 4, 27, and 108, respectively. The *minimum* average relative robustness is 0.034098 (dataset 4), i.e. on average 96.59% of the timetables produced by INALT at 1000th generation on dataset 4 are similar. As in CSALT, dataset 4 has acquired the longest average CPU time.

Table 5: Summarized Results on Schaerf Lecture Datasets using NSALT

Dataset	No. of Lectures	No. of Rooms	No. of Timeslots	Best Fitness	Average Fitness	Ave. Relative Robustness	Ave. CPU Time (seconds)
1	207	12	20	232	261.2	0.033859	420.4s
2	223	12	20	22	27.2	0.036472	80.8s
3	252	13	20	67	71.6	0.015961	145.4s
4	250	10+2	25	139	157.2	0.019271	1082.6s

The best fitness for dataset 1 is 232; i.e., 200 students without seats, 2 courses scheduled less than the specified minimum days, and 11 free gaps between lectures of the same curriculum on the same day. For datasets 2, 3, and 4, the best fitness are 22, 67, and 139, respectively. The *minimum* average relative robustness is 0.015961 (dataset 3), i.e. on average 98.4% of the timetables produced by INALT at 1000th generation on dataset 3 are similar. As in CSALT and INALT, dataset 4 has acquired the longest average CPU time.

(b) Comparing Immune-based Algorithms on Schaerf Lecture Datasets

Table 6 summarizes the experimental results on Schaerf datasets produced by the three algorithms (from Tables 3, 4 and 5).

Table 6: Comparing Immune-based Algorithms on Schaerf Lecture Datasets

Dataset	*Fitness Values*					
	CSALT		*INALT*		*NSALT*	
	Best Fitness	*Ave. Relative Robustness*	*Best Fitness*	*Ave. Relative Robustness*	*Best Fitness*	*Ave. Relative Robustness*
	Ave Fitness	*Ave CPU*	*Ave Fitness*	*Ave CPU*	*Ave Fitness*	*Ave CPU*
1	244	0.031798	256	0.148406	232	0.033859
	256.2	375.0s	264.4	529.0s	261.2	420.4s
2	25	0.023976	4	0.153383	22	0.036472
	49.6	112.0s	7.8	204.2s	27.2	80.8s
3	35	0.013898	27	0.159630	67	0.015961
	59.6	151.2s	37.2	242.0s	71.6	145.4s
4	129	0.013760	108	0.034098	139	0.019271
	145.8	4028.6s	119.6	9185.0s	157.2	1082.6s

For the *best fitness*, INALT has achieved the first position in *three* datasets (2, 3, and 4), and NSALT in *one* dataset (1). For the *average fitness*, CSALT has achieved the first position in *one* dataset (1), and INALT in *three* datasets (2, 3, and 4). It may be concluded that, for Schaerf lecture timetabling datasets, INALT is more effective than CSALT and NSALT. However, a further investigation is required to compare the average fitness of the three algorithms. For the *average relative robustness*, CSALT has the minimum values in all datasets; hence, CSALT timetables are more robust (large similarity) compared to INALT and NSALT. Finally, for the *average CPU time*, INALT has acquired the longest times on all datasets.

(c) Tests of Hypotheses: Comparing the Averages of the Fitness Values

Now the *two-tailed small-sample t-tests* are applied to further compare the averages of the fitness values, as described in Section 3.3. Three separate tests are considered for each dataset since there are three averages (three algorithms). The sample size is n = 5 (five best trials). For each test, H_0: 'two averages are equal', and H_1: 'two averages are not equal'. The test statistic t and the degrees of freedom υ are calculated using equations (3) and (4). The required *t*-tests (at 5% significance level) and the results of the tests are summarized in Table 7.

Table 7: Tests of Hypotheses (Schaerf Datasets)

Dataset	*CSALT vs. INALT*							
	$\overline{X}_1$	$\overline{X}_2$	s_1^2	s_2^2	υ	*t*-statistic	$t_{0.25,\nu}$	H_0
1	256.2	264.4	65.2	100.8	8	-1.423	2.306	not reject
2	49.6	7.8	231.3	23.2	5	5.859	2.571	*reject*
3	59.6	37.2	208.3	39.7	5	3.181	2.571	*reject*
4	145.8	119.6	201.7	104.3	7	3.349	2.365	*reject*
Dataset	*CSALT vs. NSALT*							
	$\overline{X}_1$	$\overline{X}_2$	s_1^2	s_2^2	υ	*t*-statistic	$t_{0.25,\nu}$	H_0
1	256.2	261.2	65.2	286.7	6	-0.596	2.447	not reject
2	49.6	27.2	231.3	22.7	5	3.143	2.571	*reject*
3	59.6	71.6	208.3	14.8	5	-1.796	2.571	not reject
4	145.8	157.2	201.7	244.7	8	-1.207	2.306	not reject
Dataset	*INALT vs. NSALT*							
	$\overline{X}_1$	$\overline{X}_2$	s_1^2	s_2^2	υ	*t*-statistic	$t_{0.25,\nu}$	H_0
1	264.4	261.2	100.8	286.7	7	0.363	2.365	not reject
2	7.8	27.2	23.2	22.7	8	-6.403	2.306	*reject*
3	37.2	71.6	39.7	14.8	7	-10.419	2.365	*reject*
4	119.6	157.2	104.3	244.7	7	-4.500	2.365	*reject*

From the 12 tests of hypotheses, INALT is more effective than CSALT and NSALT on *three* datasets (2, 3, and 4), and CSALT is better than NSALT on only one dataset (2). Therefore, it may be concluded that, INALT is the most effective immune-based algorithm on Schaerf lecture datasets.

(d) Comparing Immune-based Algorithms on Lecture Datasets with Other Methods

The published results by Di Gaspero and Schaerf [9] were considered to assess the effectiveness of the immune-based algorithms. They investigated the use of local search techniques based on various combinations of neighborhood functions and applied it to Schaerf datasets. A number of plain Multi-Neighborhood Hill Climbing and Tabu Search algorithms, and a number of Multi-Neighborhood (Run and Kick) Hill-Climbing and Tabu Search algorithms, have been implemented using two basic neighborhood structures *timeslot* and *room*. The best results of the *four multi-neighborhood algorithms* on four Schaerf datasets are summarized and compared with the best fitness values produced by the three immune-based algorithms in Table 8.

Table 8: Comparing Immune-based Algorithms with Other Solution Methods

Dataset	*Best Fitness Values*						
	Immune-based Algorithms			Di Gaspero & Schaerf [9]			
	CSALT	INALT	NSALT	MN-HC	MN-TS	MN-HC-RK	MN-TS-RK
1	244	256	232	285	238	200	208
2	25	4	22	18	35	17	13
3	35	27	67	72	98	55	71
4	129	108	139	140	150	113	78

[MN-HC: plain Multi-Neighborhood Hill Climbing algorithms]
[MN-TS: plain Multi-Neighborhood Tabu Search algorithms]
[MN-HC-RK: Multi-Neighborhood Hill Climbing + Kick algorithms]
[MN-TS-RK: Multi-Neighborhood Tabu Search + Kick algorithms]

From Table 8, the immune-based algorithms have achieved the first position in *two* datasets (2 and 3) compared to Di Gaspero and Schaerf [9], and the following may be concluded. For dataset 1, all immune-based algorithms are better than MN-HC, and NSALT is better than MN-TS; for dataset 2, all immune-based algorithms are better than MN-TS, and INALT is better than all multi-neighborhood algorithms; for dataset 3, both CSALT and INALT are better than all multi-neighborhood algorithms, and NSALT is better than MN-HC, MN-TS and MN-TS-RK; and finally for dataset 4, all immune-based algorithms are better than MN-HC and MN-TS, and INALT is better than MN-HC-RK. Hence, the immune-based algorithms are capable of producing good quality lecture timetables as good as other methods.

6. Conclusion

This paper has presented and compared three immune-based algorithms for lecture timetabling (CSALT, INALT and NSALT). The experimental results on four Schaerf lecture datasets have significantly shown that the three algorithms are good timetabling algorithms. The algorithms have successfully produced good quality (low fitness) lecture timetables, and hence can be applied to solve various lecture timetabling problems. The three algorithms have also successfully applied a *constraint relaxation* approach on all datasets.

Based on the fitness values, INALT is more effective than CSALT and NSALT on Schaerf datasets. The *eight* tests of hypotheses on INALT, at 5% level, have significantly shown that *six* tests favored INALT. The values of *relative robustness* have shown that the CSALT lecture timetables are more robust (large similarity) compared to INALT and NSALT. The CPU times recorded on all algorithms have revealed that INALT is the slowest algorithm. The comparison with published results [9] has significantly shown that the immune-based algorithms are capable of producing good quality lecture timetables as good as other methods such as hill-climbing and tabu search algorithms.

Even though immune-based algorithms are new in timetabling, the results produced are comparable with the well-established timetabling algorithms such as metaheuristics. The reproduction process in an immune-based algorithm does not involve a *crossover* operator as in evolutionary algorithms; instead, the main reproduction operator is *cloning*. The *power of cloning* is the vital strength that contributes to the success of the three algorithms. *Cloning and mutation* creates high similarity timetables, while *crossover and mutation* creates different timetables (small similarity).

All immune-based algorithms show great promise in the area of educational timetabling, particularly in its ability to consider, solve and optimize different lecture timetabling problems. The algorithms can handle the hard and soft constraints very well. The results have shown that the three algorithms can successfully be applied to solve and optimize various lecture timetabling problems. These algorithms may be accepted as new members of evolutionary algorithms (EAs) for timetabling. Each algorithm has all the steps involved in an EA (reproduction, genetic variation, affinity and selection). For future work, these immune-based algorithms will be employed to other domains of timetabling such as transport (driver timetabling) and healthcare institutions (nurse timetabling).

Acknowledgment

The work presented in this paper was part of my PhD thesis. Therefore, I would like to express my deep and sincere gratitude to my academic supervisors, Associate Professor Dr Ahamad Tajudin Khader and Dr Adli Mustafa, for their wise counseling, valuable advice, motivation, help and guidance throughout the duration of my doctorate study at Universiti Sains Malaysia, Penang. Without their support and cooperation, I could not have completed the study.

References

[1] Campelo, F., Guimarães, F.G., Igarashi, H., Ramírez, J.A., and Noguchi, S., A Modified Immune Network Algorithm for Multimodal Electromagnetic Problems, *IEEE Transactions on Magnetics*, Volume 42(4), pp. 1111-1114, 2006.

[2] Carlos, A., Coello, C., Rivera, D.C., and Cortes, N.C., Use of an Artificial Immune System for Job Shop Scheduling, *Artificial Immune Systems (ICARIS'2003), LNCS 2787*, Springer-Verlag, pp. 1-10, 2003.

[3] Coello, C., Rivera, D.C., and Cortes, N.C., Job Shop Scheduling using the Clonal Selection Principle, *Adaptive Computing in Design and Manufacture VI*, Springer-Verlag, pp. 113-124, 2004.

[4] de Castro, L.N., Immune, Swarm, and Evolutionary Algorithms, Part I: Basic Models, *Proceedings of the International Conference on Neural Information Processing (ICONIP2002)*, Workshop on Artificial Immune Systems, Vol. 3, pp. 1464-1468, 2002.

[5] de Castro, L.N., and Timmis, J., *Artificial Immune Systems: A New Computational Intelligence Approach*. Springer-Verlag, 2002.

[6] de Castro, L.N., and Timmis, J., An artificial immune network for multimodal, *Proceedings of the Congress on Evolutionary Computation 2002 (CEC'02)*, Volume 1, pp. 699-704, 2002.

[7] de Castro, L.N., and Von Zuben, F.J., *Artificial Immune Systems: Part I - Basic Theory and Applications*, Technical Report 1, RT-DCA 01/99, State University of Campinas, Brazil, 1999.

[8] de Castro, L.N., and Von Zuben, F.J., The Clonal Selection Algorithm with Engineering Applications, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO-2000)*, Artificial Immune Systems Workshop, pp. 36-37, 2000.

[9] Di Gaspero, L., and Schaerf, A., Multi-neighbourhood local search with application to course timetabling, *Practice and Theory of Automated Timetabling IV, LNCS 2740*, Springer-Verlag, pp. 263-278, 2003.

[10] Doyen, A., Engin, O., and Ozkan, C., A New Artificial Immune System Approach to solve Permutation Flow Shop Scheduling Problems, *Proceedings of the International XII Turkish Symposium on Artificial Intelligence and Neural Networks (TAINN 2003)*, pp. 1-11, 2003.

[11] Eberhart, R.C., and Shui, Y., *Computational Intelligence: Concepts to Implementations*, Elsevier/ Morgan Kaufmann Publishers, 2007.

[12] Farmer, J.D., Packard, N.H., and Perelson, A.S., The immune system, adaptation, and machine learning, *Physica 22D*, pp. 182-204, 1986.

[13] Forrest, S., Perelson, A.S., Allen, L., and Cherukuri, R., Self-nonself discrimination in a computer, *Proceedings of the IEEE Symposium on Research in Security and Privacy*, pp. 202-212, 1994.

[14] Fukuda, T., Mori, K., and Tsukiama, M., Parallel Search for Multi-Modal Function Optimization with Diversity and Learning of Immune Algorithm, *Artificial Immune Systems and Their Applications*, Springer-Verlag, pp. 210-220, 1999.

[15] Goldberg, D.E., Sizing Populations for Serial and Parallel Genetic Algorithms, *Genetic Algorithms*, IEEE Computer Society Press, pp. 20-29, 1989.

[16] Gonzalez, F., Dasgupta, D., and Kozma, R., Combining negative selection and classification techniques for anomaly detection, *Proceedings of the Congress on Evolutionary Computation (CEC'02)*, pp. 705-710, 2002.

[17] Hart, E., *Immunology as a Metaphor for Computational Information Processing: Fact or Fiction?* PhD Thesis, University of Edinburgh, 2002.

[18] Hart, E., and Ross, P., An Immune System Approach to Scheduling in Changing Environments, *Proceedings Genetic and Evolutionary Computation Conference (GECCO-99)*, Morgan Kaufmann, pp. 1559-1565, 1999.

[19] Hart, E., Ross, P., and Nelson, J., Producing robust schedules via an artificial immune system, *Proceedings of the International Conference on Evolutionary Computing (ICEC'98)*, IEEE Press, pp. 464-469, 1998.

[20] Ishiguro, A., Kondo, T., Watanabe, Y., Shirai, Y., and Ichikawa, Y., Emergent Construction of Artificial Immune Networks for Autonomous Mobile Robots, *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics 1997 (SMC'97)*, pp. 1222-1228, 1997.

[21] Jerne, N.K., Towards a Network Theory of the Immune System, *Annals of Immunology*, Volume 125, pp. 373-389, 1974.

[22] Kawata, Y., Morikawa, K., Takahashi, K., and Nakamura, N., Robustness Optimization of the Minimum Makespan Schedules in a Job Shop, *International Journal Manufacturing Technology and Management*, Volume 5(1/2), pp. 1-9, 2003.

[23] Koljonen, J., and Alander, J.T., Effects of population size and relative elitism on optimization speed and reliability of genetic algorithms, *Proceedings of the Ninth Scandinavian Conference on Artificial Intelligence 2006 (SCAI2006)*, pp. 54–60, 2006.

[24] Walker, J.H., and Garrett, S.M., Dynamic Function Optimization: Comparing the Performance of Clonal Selection and Evolution Strategies, *Artificial Immune Systems*, *LNCS 2787*, Springer-Verlag, pp. 273-284, 2003.

DEVELOPING A CONCEPTUAL FRAMEWORK OF KNOWLEDGE TRANSFER IN MALAYSIA E-GOVERNMENT IT OUTSOURCING: AN INTEGRATION WITH TRANSACTIVE MEMORY SYSTEM

Nor Aziati Abdul Hamid[1]
Department of Technology Management, Faculty of Technology Management, Business and Entrepreneurship, Universiti Tun Hussein Onn Malaysia (UTHM), 86400 Parit Raja, Batu Pahat, Johor, Malaysia aziati@uithm.edu.my

Juhana Salim[2]
Head of Information Science Program, Faculty of Technology and Information Science, Universiti Kebangsaan Malaysia (UKM), 43600 Bangi, Selangor, Malaysia
js@ftsm.ukm.my

Abstract. Knowledge transfer has attracted much attention to researchers and practitioners in recent years since knowledge transfer has been considered as a critical determinant of an organization's capacity to confer sustainable competitive advantage. Despite extensive research on knowledge transfer issues, there is a dearth of research that has explicitly focused on the role of transactive memory in enabling intra-organizational knowledge transfer in information technology (IT) outsourcing context, particularly e-government IT outsourcing. Although the information systems literature has recently acknowledged the role of transactive memory plays in improving knowledge processes, most of the research is still in the basic concept of transactive memory which is emphasized more on the individual level of analysis or rather in the small group of people. Besides, most of related research was done in the lab based on the physical, virtual task or memory recall tasks. None of empirical work has been done in integrating TMS in outsourcing context since most researchers used interpretive approach. Therefore, this paper attempts to fill this gap by applying positivist approach through operalization of identified factors that significantly give positive impact towards knowledge transfer between the vendors to the Malaysian Public Agencies as the client.. Drawing on several theoretical streams, this paper will propose an integrated conceptual framework of inter-organizational knowledge transfer with and integration of transactive memory system to facilitate knowledge transfer process between organizations which further can be used for research enhancement.

Keywords: Knowledge Management, Knowledge Transfer, Organizational Learning, Transactive Memory System (TMS), Information Technology Outsourcing (ITO).

1. Introduction

Knowledge Management (KM) has been historically influenced by research undertaken across broad range of disciplines. These disciplines include sociology, psychology and philosophy. Until now, research in KM has been extended through various areas such as strategic management, information system, organizational learning, artificial intelligent and other more. Among those parent disciplines, organizational learning is the closets 'cousin' to KM *"with KM and organizational learning being considered two sides of the coin"* [1]. Transfer of knowledge is critical to knowledge-intensive project like IT outsourcing. However, the transfer of knowledge requires continuous organizational learning and the knowledge is being organized to enable knowledge retention capacity for future knowledge utilization. Knowledge is considered as tangible asset to organization. Tangible assets tend to depreciate in value when it is utilized. Knowledge grows substantially when it is fully utilized and depreciates or stagnant when it is not used. The organization needs to acquire the knowledge, learn, apply and reinvent the knowledge to make it suitable with the organization climate. Indeed, knowledge is of limited value if it is not shared and transferred throughout an organization. Thus, interest has increased in the phenomenon of how the firms create, retain, and transfer knowledge.

In the case of Malaysia, Malaysian Administrative and Modernization Planning Unit [2] has created a "knowledge bank" structure in the public sector ICT framework to facilitate the sharing of knowledge and experience by capturing information across all Government agencies. This framework will create a structured and systematic transfer and utilization of knowledge generated. For the initial stage, several sets of databases has been identified by MAMPU such as economic intelligence, security intelligence, R&D and Government statistics to create the knowledge bank. This initial project is implemented at four ministries; Finance, Health, Works and Education Thus, each ministry must develop their own knowledge bank with back end architecture that can integrate with other stated ministries. This project was initiated to address a high number of complaints regarding public services. There are many factors contribute to poor service delivery in the public sectors and one of them is low level of information and knowledge sharing among government agencies [3]. Although there is an increment in term of percentage complaints solved, the success story of knowledge bank implementation reported by the scholars and how it can facilitate the knowledge transfer process is scarce. Furthermore, the knowledge bank focused more on the internal knowledge repositories among the public agencies without the absence of private agencies [2]. Since the government have been aggressively promoting the Shared Service and Outsourcing (SSO) industry, which undertake a full consideration of public-private partnership in supporting government transformation, it is crucial to consider a suitable framework of knowledge bank that could support and facilitate transferring process during the partnership.

Past researchers have suggested various organizational, human-related and IS-based mechanisms for improving knowledge transfer processes within and between organizations. Recent research has starts to integrate the concept of individual’s mental memory towards organization. Organization by itself is a combination of various stages of memories ranging from internal memories until external memories (e.g. stakeholders/shareholders). Therefore, this paper attempts to provide a better understanding of the phenomenon of knowledge transfer in IT outsourcing and how the transferring process may be bridged by applying organizational memory concept with existing identified factors during IT outsourcing project execution in Malaysia government setting. Next section will discuss the context of this study with supports from relevant literature to develop underpinning theories in section 3. In section 4, we suggest a conceptual framework of knowledge transfer in IT outsourcing with proposed hypotheses for each concept. We conclude the paper by some final remarks in section 5.

2. Research Background

2.1 Knowledge Transfer and IT Outsourcing

Knowledge transfer have been defined by most scholars as a dyadic exchange between individuals, groups or organizations in which a recipient can understand, learn and apply knowledge transmitted from a source [4][5][6]. A thorough review of literature reveals that many authors and researchers have failed to provide a clear cut definition for KT and at the same times use the term "knowledge sharing (KS)" and "knowledge transfer (KT)" interchangeably. However, recent scholars' works have made a distinction line between these two terms. Knowledge sharing primarily concerned with the individual's view while knowledge transfer concentrates more on the organizational view [7]. KS only takes the activities of giving or contributing, and is included under sub process of knowledge transfer. Furthermore, Kumar and Ganesh [8] asserted that KS does not include the receiving and reuse aspect of transfer. KT should involve active communication between two parties or active consultation for each other in order to learn what they both know. In a simple connotation, "people share knowledge" whereas "organizations transfer knowledge".

Some researchers have been arguing of knowledge transfer concept since knowledge resides in employees (human components of organization), task and interrelationship, tools and technology (software and hardware) and network coordination (internal or external network coordination). There is no simplest way to transfer knowledge from a brain of a human to another brain perfectly and easily like transferring files form one computer to another. Hence, the nucleus of knowledge transfer process is the knowledge receiver. The knowledge receiver must have capability to learn, to understand and to know for applying in right circumstance. In line with that, all knowledge transfer mechanism incorporates social interaction either from direct interaction or virtual interaction. Ambos and Ambos [9] identified two mechanisms; (1) by personal coordination mechanism such as personnel motion, training, jobs rotation [10], interactions with suppliers and customers [11], community of practices and post-project reviews [12], (2) by technology based coordination mechanism such as collaboration software, distributed learning and business intelligence system. Most of Malaysia organizations are actually practicing knowledge transfer using mechanism like staff training, observation of experts, routines, meetings, standard operating procedures, manuals and databases where most of transferring knowledge process is the implication of strategic alliances, joint ventures, mergers and acquisitions. KT especially through strategic alliances has become a shot gun approach for a firm to acquire knowledge that it could not easily develop within its confines. One of the strategic alliances practices in Malaysia is through IT outsourcing.

During IT outsourcing partnership, client and vendor can develop two forms of knowledge transfer in terms of a reciprocal learning [13]; 1) the partners can obtain from each other technical knowledge and know-how, 2) they can learn from each other management and business skills that individually they are lacking. Both the service receiver and provider should have a shared vision and goals for partnership as well as a belief that their partners will not act opportunistically [14]. Knowledge transferring or sharing throughout the IT outsourcing progress management should be given more attention for both sides. One side, vendors can transfer their IT special knowledge to clients, which helps client to improve their IT function; on the other aspect, clients also transfer their business knowledge to vendors, which will improve vendor's capability of understanding and implementing. Unfortunately, it appears that public sector organizations in developing countries especially Malaysia, have not received much attention in the research literature covering knowledge transfer especially in IT outsourcing. Most of the studies concentrate on the general knowledge management implementation or readiness at public agencies [15], Malaysian SME industries [16], aerospace industry [17], bank [18], telecommunication industry [19], higher education [20] to cite a few. There is only one work recently done by Mohamed et al. [21] focusing on knowledge transfer success factors in Malaysia setting. From the success factors the authors developed a theoretical framework for future work. Apparently, those researches never address the need of organizational learning context for an effective knowledge transfer. Therefore, it is crucial for this study to be taken and significantly give an insight and better understanding of the knowledge transfer processes in ITO.

2.2 Malaysia E-government IT Outsourcing Initiatives

In today's world, governments are increasingly under pressure for more profound change in structure and strategies to meet the requirements of contemporary society. Government needs to become more partnership-based, results-oriented, integrated, and externally focused. Therefore, government starts to serve their citizen thru electronic application. Malaysian government has starts their initiative in transforming their service delivery by launching seven flagship of e-government with the development of Multimedia Super Corridor (MSC) has become a jump-starts of all current transformation. In order to focus more on servicing citizen, e-government outsourcing has become an important measure to reduce the pressures from cost, technical, as well as personnel. E-government Outsourcing in the Malaysia public sector has become an accepted management practice. Yang et al. [22] classified e-government outsourcing into two types; (i) system construction outsourcing (project in nature) and (ii) maintenance outsourcing (process in nature). Usually e-government outsourcing project will involve two or more vendors working together for one particular project. The relatively high complexity, high uncertainty, and high risk of large e-government service projects favour a partnership approach. This government (clients)-private (vendors) partnership make the knowledge transferring process more problematic due to differences in the development and implementation of IS across sectors.

According to a joint publication by Outsourcing Malaysia and Value Notes published in August 2009, revenues from the Malaysian ITO industry are expected to touch $1.1 billion in 2009. The industry is expected to grow at a CAGR of 15% to reach $1.9 billion by 2013. Currently, ITO services in Malaysia have a greater share of the overall outsourcing market, followed by Business Process Outsourcing (BPO) services; while knowledge services outsourcing is still in its nascent stage, has a smaller share. The interest in outsourcing is still growing especially among players in the banking (e.g.: CIMB & Maybank), airline (Malaysia Airline System), manufacturing, healthcare, and government sectors. IT outsourcing has been identified as one of the main ways to address some demanding challenges faced by government. The shortage of IT expert and the difficulty of attracting and retaining the right IT talent ranked as the number one barrier that fuel the Malaysian government decision to outsource. Current e-government IT outsourcing activities in Malaysia are data entry, ICT hardware maintenance, network management service, web-hosting management and development and application system maintenance [2]. However, there is a trend for government and public agencies to shift to more interactive service delivery which are citizen-centered and based on networks and partnership between public, private and NGO and between levels of government. The use of application providers by government can help meet increasing e-government service demands by citizen and business alike.

Currently, Malaysian government has been practicing three types of IT outsourcing model for e-government application namely [2]; (1) BOT (*Build, Operate, Transfer*), (2) BOO (*Build, Operate, Own*) and (3) Contract Services. For *BOT* approach the provider/vendor need to develop the application according to the agencies requirement and manage the system operation for a certain time as stated in the contract. After the contract terminate, the vendor will hand over the application to the agencies that owned the project. Example applications for BOT approach that have been implemented are e-procurement (e-perolehan) own by Ministry of Finance (MOF) and The Electronic Budget Planning and Control System (e-SPKB) own by National Accountant Department (ANM). In contrast with *BOO* outsourcing approach, the vendor will provide and manage the ICT service without hand in back to the agencies. The ownership of the services is still under vendor supervision. The last outsourcing approach is *contract basis service*. For this approach, the owner agency will give a contract to the vendor to develop/maintain the whole ICT devices but the ownership of the device belongs to the agencies not the provider.

It shows that Malaysian government has massively outsourced many e-government applications but scarce researches have focused on knowledge transfer processes in the outsourcing projects particularly for Malaysia environment. Although most of the success factors for ITO were rigorously considered based on principles and findings from previous research, which are frequently referred to [23], there are still some project that is not fully satisfied by the stakeholders or do not meet stated performance objectives [24]. Report from egov4dev.org (2009) has shown that e-government project failed because there is no lesson learned since knowledge about the failure was not captured, transferred or applied. As a result, mistakes were wastefully repeated. This claimed was also supported by Giannakis [25] which examined the importance of knowledge transfer towards vendor's development that can create added value to the organizations. Giannakis [25] asserts that the failure of many initiatives revealed a twofold problem: first there is great difficulty in the generation and transformation of knowledge into organizational action and subsequently and even greater difficulty in the transfer of knowledge to partners. In addition, the acquired application may not be customized enough to effectively streamline or transform the business process. Moreover, this relates to the criticism that the vendors have limited understanding of the clients' business process [26]. IT outsourcing involves integrating and coordinating knowledge from many individuals of different disciplines and backgrounds, with varied experiences and expectations, located in different parts of the organization. Thus, both client and vendor should able to identify types of knowledge that is needed to be transferred during project execution, what mechanisms are appropriate and how the transferred knowledge can be retain in the organization for learning purposes and future use. To address the issues, we have drawn our research from two popular theories in Knowledge Management field as well as outsourcing field.

3. Theoretical Lens

According to Benedikt and Frank [27], the popular theories being used in ITO research is the economic theory (e.g. Transaction Cost Theory & Agency Theory), followed by sociology theory (e.g. Relational Exchange Theory & Social Exchange Theory) and lastly strategic management theory (e.g. Resource-Based Theory, Resource Dependence Theory). From the researcher literature review, for the past five years research in ITO and knowledge transfer, most researchers used multiple theoretical approaches rather than single theoretical approach. The most dominated theory behind the knowledge management activities in ITO project was two popular models; Resource-Based View Theory (RBV) and Knowledge-Based View Theory (KBV). From a sourcing perspective, RBV theorists have traditionally maintained that firms should not outsource any business function or activity that contributes to building and maintaining competitive advantage. According to this two theories postulate by Barney [28] and Wernerfelt [29], firms that established connections with external firms through mechanisms such as outsourcing run the risk of transferring vital knowledge and resources by engaging in sourcing partnerships. Other potential negative sourcing outcomes include creating competitors via vertical integration of sourcing partners and losing vital internal knowledge and resources by engaging in sourcing relationships with external partners. As a result, RBV called for a protectionist stance regarding outsourcing, recommending that firms should only outsource support functions that do not directly contribute to the firm's value added and competitive advantage generating mechanisms.

From a more proactive perspective, RBV and KBV tenets denote that firms may engage in outsourcing as a means of identifying, exploring, and transferring knowledge and resources from external sourcing partners to internal control. KBV proposes that IT outsourcing is a way to utilize vendor's professional knowledge and skills [30]. Although the knowledge-based view emphasizes the unique knowledge of the client firm, IT projects needs an integration of mix experience and new knowledge from the vendor. Client and vendor firms can create shared understanding from a successful exploration of specialized external knowledge. The exploration of external knowledge in IT outsourcing needs a knowledge integration of client domain knowledge and vendor technical knowledge during the development process. Without such integration, the unique knowledge of the client firm cannot be successfully leveraged in the outsourced custom-software development process. Consequently, IT outsourcing can be viewed as a boundary crossing mechanism through which firms can use sourcing relationships to gain access to resources critical to the firm's competitive advantage development or maintenance [31]. In such cases, client establishes a short-term relationship with an established outsourcing partner with the intent of transferring knowledge, human capital, and technologies from the client to the vendor. Additionally, Combs and Crook [31] asserts, mechanisms emphasized in outsourcing strategy can range from the (i) transferring of knowledge to help develop internal capabilities, (ii) by the hiring an experts personnel from the sourcing firm to build up internal capabilities for the partner, (iii) by the outright acquisition of the sourcing firm to internalize capabilities previously existing externally and lastly (iv) by aligning client's needs with vendor possessing complementary resources and capabilities. In such cases, outsourcing partners may provide the combination of complementary knowledge bases with a lack of direct competition that can fuel innovation of a new application/technology and service development. Hence, RBV and KBV perspectives provide valuable insights for the business rationale of IT outsourcing practices. Many researchers have found them useful in explaining specific aspects of outsourcing decisions, processes and outcomes using KBV and RBV theoretical lens. Thus, many researchers (e.g. [32], [33]) have placed these two theories as the theoretical lens to the KT model or their framework specifically for ITO environment.

3.1 Knowledge Transfer Model

King et al. [4] appointed two important element in developing effective organizational knowledge; (i) communication and (ii) information processing. There are three models dominated within the knowledge transfer area. Most of the existing KT models were rooted from communication model, group information processing model and knowledge creation model. Communication based model was elucidated by Schramm [34] and later being improvised by Jacobson [35] while the second is based from Hinsz's [36] model. The third one is based from Nonaka's [37] knowledge creation model. Within the communication-based approach, the transfer of knowledge is regarded as a message encoded in a medium by a sender to a recipient in a given context. Schramm's [34] communication model initially consisted of three elements; (i) Sender, (ii) Recipient and (iii) Message. The receiver becomes the "recipient" or "user", since it is the subject who learns or acquires knowledge (not simply the message receiver) whereas; the "sender" is the knowledge holder. The message becomes the "object", as it can be produced by complex knowledge. Scharmm's [38] later enhanced the model by including Media. Media is the channels used to communicate the message, palliate its passage, and enhance its chances of completing a communicative act. Scharmm's [38] model becomes the most referred basic model in many knowledge transfer framework. Subsequently, Jacobson [35] improvised the basic model developed by Schramm's [38] by considers six factors: Knowledge source, Message, Knowledge receiver, Channel, Feedback and Environment or Organizational context.

Apart from viewing KT from communication lens, scholars started to integrate the communication model with group information processing model to enhance the existing KT model. In order for the organization to learn something, the members need to process the data or information that they got to better suit the organization. Hiss et al., [36] has postulated three components in the information- processing model: *encoding* (i.e. forming knowledge representations through interpretation, evaluation and transformation), *storing* (i.e. entering representations in the memory system), and *retrieval* (i.e. accessing and using representations from the memory system). This concept is closed to human cognitive system. Later, Gibson [39] starts to improvised Hinsz's [36] model by expanding the information processing into four stages; accumulation, interaction, examination and accommodation. However, Gibson's model is applicable if the accumulated knowledge is highly ambiguous and the processing does not occur in a linear time order. The main similarity between these two models is the need of social interaction along each phase.

From Nonaka and Takeuchi [40] framework of knowledge generation, the transfer of knowledge is seen as the creation of knowledge through four modes of knowledge conversion of explicit and implicit forms of knowledge: externalization (from implicit to explicit), combination (from explicit to explicit), socialization (from implicit to implicit) and internalization (from explicit to implicit). Nonaka and Takeuchi [40] visualized the knowledge conversion process as cyclic process and happen mainly through informal networks of relations in the organization starting from the individual level, then moves up to the group (collective) level and eventually to the organizational level. However, according to Sherif and Xing [41], Nonaka's [37] model does not describe how to initiate the macro level process for individuals or groups to manage the knowledge and to be innovative. Gourly [42] further claimed that Nonaka only proposed two modes of knowledge creation; internalization and externalization, whereas; socialization and combination are modes of knowledge transfer. Based from the discussed constraint, Curseu [43] develops an integrative cognitive architecture model for groups with the combination of three subsystem; selection subsystem, memory subsystem and communication subsystem. Curseu [43] claimed that the comprehensive group information processing models should consist of communication based view, knowledge creation based and memory based system. The proposed information processing model can be integrated with communication model, Gibson's [39] model, Nonaka's [37] Model and Transactive Memory theory. This model is suitable at the organizational level unit of analysis for example this model appropriate for distributed group members and virtual project team. However, the group members must be anonymous.

Besides the three basic models as the basis to the KT model developed by past researchers, scholars have also embodied KT antecedents and consequences in the model. Prior studies have investigated the role of knowledge characteristics, such as ambiguity and complexity, in determining knowledge [10]. Other studies have examined sender-receiver characteristics; such as absorptive capacity and learning capability [44],[45] or organization context [46],[47]. Inspite of that, current trends in knowledge transfer research have also comprised project nature [5],[33] factors in developing the model since most of the transferring process occurred during the project execution or alliances. Table I summarized a few KT components that being derived from the past research. These components have been reviewed by most of the scholars in KT research and significantly gives effect on KT process in ITO.

In spite of all factors discussed above, organization information system is claimed to be an effective tools to support knowledge transfer process. However, most of the organizational knowledge is based on the information stored in legacy information systems which have been developed in an isolated way [48]. Therefore, such information can be inconsistent, redundant and difficult to retrieve and link. The information that ends up in the most organizational information system has a poor structure (e.g., PDF documents), which makes the system unmanageable and chaotic, limiting the possibility to deal with other system requirements, such as information privacy and fast and flexible retrieval methods [48]. It was suggested that the organizational information memory system should have the capability to provide an experts database with points of contact on various topics [49], support both formal and informal knowledge besides the automatic privacy mechanism [48]. Hence, recent scholars have connected organizational information memory system (OMIS) with the Transactive Memory System (TMS) to facilitate the interaction of organizational knowledge [50][51][52].

Table 1: Knowledge Transfer Components

Components	Characteristics	Authors
Source	Disseminative Capacity Reliability Credibility Willingness to share	[10][33][45][47]
Recipient	Absorptive Capacity Motivation Learning intent Retentive Capacity	[45][66][67]
Knowledge	Knowledge Ambiguity Stickiness Complexity Tacitness	[45][66][68]
Organizational	Organizational Culture Personnel Movement Community of Practices Management Practices Organizational Structure Organizational Learning	[69][70][71][72]

	Strategy	
Communication	Codification Interpretation Communication Channel	[73]
Relationship	Arduous Relationship Dyadic relation Strength of ties Network density Social Similarity	[6][10][73]
Project Nature	Prior collaboration history Team size Project complexity Project phase	[33]

3.2 Organizational Memory System and Knowledge Transfer Process: The role of Transactive Memory System

Knowledge transfer (KT) is a multilevel process whereby the transfer activities not only individually held skills, but also organizationally embedded knowledge or collective knowledge. According to Narteh [53], KT process comprises four activities; knowledge conversion, knowledge routing, knowledge dissemination and knowledge application. Within these practices, effective transfer and use of organizational knowledge depends to a large extent on the organization's ability to create and manage its collective memory. The organization itself has been seen as a repository of knowledge [54]. The organization's knowledge repositories or knowledge stock are found in individual members, roles and organizational structures, standard operating procedures and practices, culture and physical layout of the workplace [53]. This collective memory is often referred to as organizational memory (OM). To support effective management of organizational memory, Stein and Zwass [55] proposed the use of information technology to accomplish four specific processes related to organizational memory: acquisition, retention, maintenance, and search and retrieval. In addition, they outline a design for an organizational memory information system (OMIS).

However, Nevo and Wand [56] argued that the proposed OMIS architecture by Stein and Zwass [55] faced several challenges. According to them, much of the knowledge in the OM is contextualized and consequently the knowledge interpreted wrongly. A second challenge regarding the locations of knowledge since OM generally resides in different types of retainers. These retainers of OM may be in dispersed location and their memories might be difficult to combine. A third problem with OM management is that knowledge is often tacit which is difficult to track and maintain in large organizational memories. A fourth problem concerns with the unpredictability of organizational knowledge. This unpredictability results in frequent changes to the contents of the OM measure of the retainer's legitimacy and reliability is required. These five problems create difficulties for members of the organization in retrieving and using knowledge that resides in OM. Therefore, to gain a better understanding of possible ways to overcome the barriers for efficient OM management, Nevo and Wand [56] proposed the concept of Transactive Memory Systems (TMS) being incorporated with OM.

One of the philosophical theories that have been embedded in the concept of organizational memory is Transactive Memory theory. Transactive Memory theory becomes Transactive Memory System when Wegner [57] started to model human memories in a concept of memory sharing in computer systems. Transactive memory is a system for encoding, storing, and retrieving information in groups [57]: it is a set of individual memory systems in combination with the communications that take place between individuals. Originally, TMS was used to describe the ways in which dyads (such as married couples) that are close to one another share knowledge and allocate responsibilities for knowing. Extending the notion of TMS beyond groups and pairs, several authors have speculated on how organizations might function as TMS with an input of information system architecture. Anand et al. [58] proposed certain forms of information systems, such as intranets, search engines, standardized concepts and vocabularies, could be used to enhance the functioning of TMS. Nevo and Wand [56] proposed directories of meta-knowledge to overcome the knowledge storage and location problems as stated before. The computerized directories of meta-memory can compensate for the lack of the group's tacit knowledge. Even so, the work on organizational TMS has been conceptual rather than empirical. There have been no descriptions of working organizational TMS in the literature.

Therefore Jackson and Klobas [51] have proposed a model of the operation of an organizational TMS. This model focused more on organizational KM codification strategy rather than personalization strategy since the aim of suggested model was to connect people with reusable codified knowledge. Jacobson and Klobas [51] have divided organizational TMS into four main activities instead of three activities postulated by Wegner [57]. The nucleus of organizational TMS is the directory or the knowledge repositories. The directory consists of metadata about people, including name, organizational role and formal group membership, work experience, areas of expertise and other information such as availability and reliability as a source of knowledge. Some of the metadata for some people in a TMS will be stored in a person's head, but other metadata can be stored externally, in a CV or expertise database, a document management or knowledge management system, on the organization's intranet or in handbooks, or in the heads of intermediaries such as managers, administrators and other colleagues who act as gatekeepers or links in a chain to the ultimate source of the knowledge. The second activity is directory maintenance. According to them, directory can be maintained by formal and informal procedures. Formal procedures might include the updating of metadata and other information in organizational information systems whereas informal procedures include discussions held alongside formal meetings or serendipitous meetings in the corridor or coffee room. The third activity is retrieving process from the directory. The directory allows knowledge to be retrieved from one's own work group(s) and from others in the organization. Much of the information retrieval from one's own group might be in the form of conversations although this retrieval might be supported by information systems that record knowledge in the form of documents. Finally, knowledge allocation would be the fourth activity evoked by Jacobson and Klobas [51]. They argued that knowledge is allocated and stored on the basis of several activities ranging from formal allocation of responsibility and transfer of knowledge among people in the organization to individual learning. This view provides a framework to guide development of a holistic TMS for a particular organization. It allows a view of what an information system might provide and what is best done (or indeed must be done) through interpersonal means.

4. Conceptual Framework

The underpinned framework for this study is derived from the in-depth study on IT/IS outsourcing, knowledge transfer, information processing literature and organizational learning. Previous research has examined a range of antecedents of organizational knowledge transfer. For this research purposes, this study included only antecedents that have been studied extensively across multiple studies and align conceptually. This enabled researcher not only to compare antecedents, but also to make sure the antecedents studied are deemed relevant by the research community. Consistent with prior literatures, the researcher classifies antecedents of inter-organizational knowledge transfer into four domains: organization memory system factors, client-related factors, vendor-related factors while project management factors as controlled variables. This paper contributes to the existing literature by examining how organization memory system can facilitate the knowledge transfer process between client and vendor involved in IT outsourcing relationship besides the other three most cited determinants. From the IT project management perspectives, organization shared cognition are able to successfully manage project interdependencies [59]. Fig. 1 illustrates the proposed conceptual framework for the study.

4.1 Variables

The dependent variable in the research framework is ‘knowledge transfer’. The operationalize definition of knowledge transfer for this research was drawn upon the communication theories, whereby transfer of knowledge is define as a method that involves two-way communication between the client and the vendor exchange and share their useful information/ skill/ competencies or routines about the project and both parties is affected by changes in recipient replication and adaptation capabilities and changes in skills/knowledge. Knowledge from this research context is organizational knowledge whereby “knowledgeable” organization can be seen through daily basis routines and the systematic structure of workflow. A vendor corresponds to the knowledge source involved in transferring knowledge or the generalized knowledge resource, whereas; client act as knowledge receiver, and the destination or the entity which receives and internalizes the knowledge content. Further, within the knowledge transfer context, the transmission element corresponds to the activities and processes, such as communication activities, through which knowledge is transferred from one entity to the other.

Meanwhile, the independent variables are measured by three domains; vendor characteristics, client characteristics and organizational memory context. Each of the domains is observed by several items that have been selected from Table 1. Researchers only take the items that empirically give significant or positive impact towards knowledge transfer. The negative impact has been eliminated to ensure the high validity and reliability of each construct. Client in this research context is the Malaysian public agencies act as the recipient of knowledge that outsourced the E-government application to the third parties. Meanwhile, vendor is conceptualize as a third-party entity act as the source of knowledge that develop, manages and distributes E-government application and solutions to public agencies. Vendor characteristics are measured by vendor credibility, willingness to share, disseminative capacity and knowledge integration. For client characteristics, researchers have chosen four measurable item; absorptive capacity, retentive capacity, conjecture and motivation. Researchers have also incorporated Transactive Memory System (TMS) as proposed by Jacobson and Klobas [51] and Oshri et al. [50]. Although the most popular measurement of TMS is elucidated by Lewis [60], by which TMS is measured by specialization, coordination and credibility, we argue that the early measurement developed by Lewis [60] is based from the activities memory recall of dyads that working together and it is most suitable of TMS form individual's perspective rather from organizational perspective. Therefore, we have extracted the main organizational routines that involves during outsourcing project as presented in Oshri et al. [50] and Kotlarsky and Oshri [61] interpretive research. Therefore, TMS is measured from the organizational project routines that encoding and updating directories, coordinating and retrieval, allocating and storing and lastly directory content.

Much of the academic research on information system project management has been done in the context of software development and maintenance in the "traditional" computing paradigm in which the majority of software projects involve the custom development of applications [33]. There is a lack of empirical investigation of the issues related to the IT outsourcing projects. Control variables in this model are derived from project management literature, but we labelled it as project nature. Thus in this research, four control variables are included in the framework: prior collaboration history, team size, project complexity and project phase.

4.2 Framework Hypotheses

The framework presented in Fig. 1 organizes knowledge transfer research based on several areas of emphasis including client characteristics, vendor characteristics, organizational memory characteristics and the nature of project. Each area of emphasis consists of related topics that we identified in our review of knowledge transfer research. The dependent variables are vendor characteristics, client characteristics and Transactive Memory System (TMS), while independent variable is knowledge transfer in IT outsourcing. Each of variables is measured by the most cited constructs and highly impact towards knowledge transfer effectiveness. Thus, we derived 15 hypotheses from each area that relates.

4.2.1 Vendor Characteristics

Many studies have examined the effect of knowledge source on knowledge transfer. The knowledge source in this research refers to the vendor that develops or provides the e-government applications or infrastructures. There are four characteristics of vendor that being measured in this study; vendor's credibility, vendor's willingness to share, vendor's disseminative capacity and vendor's capability to integrate knowledge from various units/departments. Vendor's credibility is generally defined as the extent to which a client perceives a vendor to be trustworthy and reputable [33]. Thus from the definition, the credibility concept has two dimensions: trust and reputation. Knowledge transfer researchers have indicated trust as the core ingredient in order for individuals to transfer knowledge [10]. Trust 'reflects the belief that a partner's word or promise is reliable and that a partner will fulfil its obligations in the relationship' [54]. When client credibility is high, client are likely to be more open and receptive to information from the vendor; ideas in the asset are perceived to be worthy of consideration. The knowledge conveyed is thus more likely to be seen as useful, and to influence the behaviour of the recipient [10]. The importance of a client's credibility is amplified in the context of a knowledge transfer process because this process is not amenable to enforcement by contract [62].

Besides vendor's credibility, we also measures knowledge sharing initiatives in the project. Lee ([14] and Joshi et al. [10] have showed that knowledge sharing is a major indicator of whether or not the outsourcing activity succeeds. Those studies confirms that knowledge sharing is one of the major predictors for outsourcing success because IT outsourcing posses highly valuable knowledge relating to the product development, the software development process, project management and technology in general [63]. Therefore we operationalize willingness to share as vendor attitude which vendor is willingly to provide access towards others about knowledge and his experiences. Willingness to share is operationalized based on the intensity level of vendor in doing tacit and explicit knowledge sharing with his client in ITO project. Willingness to share also relates to the vendor's disseminative capacity. Disseminative capacity refers to the vendor capacity to contextualize, format, adapt, translate and diffuse knowledge through a social or technological network and to build commitment from stakeholders [64]. In the context of IT outsourcing, individual members who control and distribute resources, information and knowledge can largely affect the performance of the whole project team [65]. The fourth constructs is knowledge integration. Knowledge integration is defined as individual members who control and distribute resources, information and knowledge can largely affect the performance of the whole project team [65]. In an IT outsourcing project, the users from the client organization communicate system requirements to the vendor's IT consultants who use their software expertise and knowledge from the users to build the system. Users then assimilate the system by making necessary changes to their work. Knowledge integration is essential since if knowledge from a particular cluster is missing or is not integrated. Therefore, we derived five hypotheses from vendor characteristics:

H1: Vendor characteristics significantly give an impact towards vendor characteristics for knowledge transfer processes in IT outsourcing at public agencies

H1a: Vendor credibility significantly gives an impact towards vendor characteristics for knowledge transfer processes in IT outsourcing at public agencies

H1b: Vendor willingness to share significantly gives an impact towards vendor characteristics for knowledge transfer processes in IT outsourcing in public agencies

H1c: Vendor disseminative capabilities significantly give an impact towards vendor characteristics for knowledge transfer processes in IT outsourcing in public agencies

H1d: Vendor knowledge integration significantly gives an impact towards vendor characteristics for knowledge transfer processes in IT outsourcing in public agencies

4.2.2 Client Characteristics

Second independent variables involved in this research are the client factors. There are four independent variables has been identified in this research that influence the process of transferring knowledge in IT outsourcing project; absorptive capacity, retentive capacity, communication competence and motivation. Most scholars stress that the studies of knowledge transfer should concern not only whether knowledge owners have a willingness to share, but also whether knowledge receivers can learn and absorb. Therefore, absorptive capacity affects vendor ability to recognize the importance and value of new knowledge, to assimilate the knowledge, and to apply it to solve the problem. We defined absorptive capacity as the ability of the client to acquire new external knowledge, assimilate or transform the knowledge into usable knowledge then apply it to business ends. The definition emerges two subsets; potential absorptive capacity and assimilation and realized absorptive capacity. The client needs to actually know their prior knowledge and their ability to valued new knowledge that they received from vendor for example through training or project maintenance. A transfer of knowledge is effective only when the knowledge transferred is retained. The ability of a recipient to institutionalize the utilization of new knowledge reflects its 'retentive' capacity [10]. According to Schwartz [7], clients retentive capacity is differs from absorptive capacity because absorptive capacity refers to an indication of initial short-term memory, whilst; retentive capacity refers to long-term memory.

Communication competence can be defined as the extent by which the client and vendor have a frequent routine of formal (in term of task-achieving issues) or informal (out of role) interaction and conversation regarding project-relevant information. The uncertainty situation in IT outsourcing, my impacts the process of knowledge transfer among clients' and vendors' that emerged the important of communication competence. On top of that, client needs the motivation to accept and absorb the new external knowledge. The motivation of the client refers to the client desire to implement the knowledge or technology being transferred. Lack of motivation in knowledge transfer will result in passiveness, feigned acceptance or implementation, hidden sabotage, intentionally slow implementation, or directly reject the practice. From the above argument, we posit five hypotheses from client variables;

H2: Client characteristics significantly give an impact towards client characteristics for knowledge transfer processes in IT outsourcing in public agencies

H2a: Client absorptive capacity significantly gives an impact towards client characteristics for knowledge transfer processes in IT outsourcing in public agencies

H2b: Client retentive capacity significantly gives an impact towards client characteristics for knowledge transfer processes in IT outsourcing in public agencies

H2c: Client communication competence significantly gives an impact towards client characteristics for knowledge transfer processes in IT outsourcing in public agencies

H2d: Client motivation significantly gives an impact towards client characteristics for knowledge transfer processes knowledge transfer in IT outsourcing in public agencies

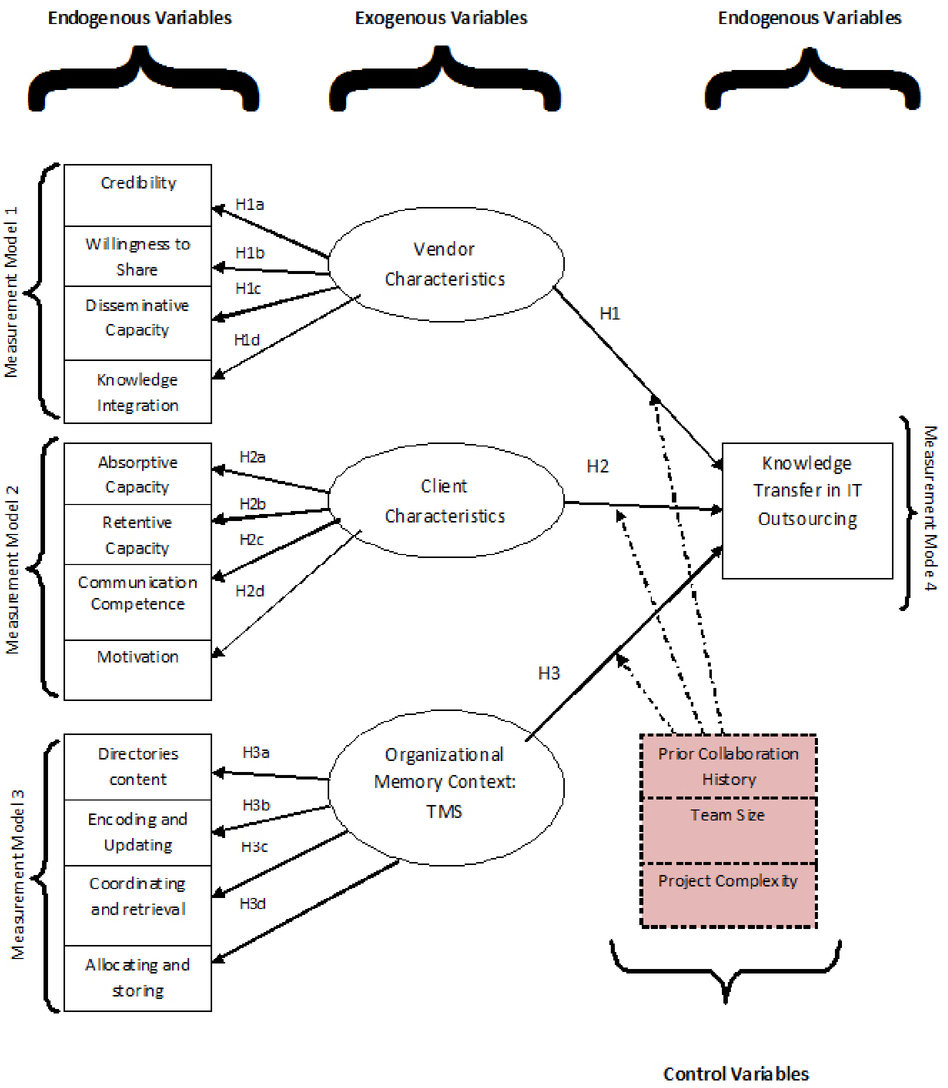


Fig. 1: A conceptual framework of knowledge transfer in e-government it outsourcing

4.2.3 Organizational Memory Context: TMS

Knowledge transfer in group encompasses various practices of managing organizational knowledge. Effective transferring and use of organizational knowledge depend on a large extent of the organization’s ability to create and manage its collective memory. This collective memory is often referred to as organizational memory (OM). In relation to IT outsourcing, such memory resides in business professionals from clients’ side and IT specialist from vendors’ side, policies, contract/agreement, and culture. These retainers of OM may be in different locations and their memories might be difficult to combine. Thus, scholars have increasingly considered the concept of the Transactive Memory as an enhancer of inter-organizational knowledge transfer [50][56] and to develop organizational knowledge memory system [51]. While the concept of transactive memory has been studied in the context of traditional organizational forms and co-located teams, little is known about the process through which a TMS in distributed teams could be created and could support knowledge transfer between remote sites like a case in IT outsourcing. We measures TMS in terms of the project routines to encoding and updating the directories, coordinating and retrieval process, allocating and storing of project information or data in the organizational memory systems. Thus, we believe that TMS will facilitate the knowledge transferring process;

H3: Transactive Memory System significantly facilitate knowledge transfer in IT outsourcing at public agencies

H3a: The relevant organizational directories content significantly facilitate knowledge transfer in IT outsourcing at public agencies

H3b: The project routines of encoding and updating project document significantly facilitate knowledge transfer in IT outsourcing at public agencies

H3c: The project routines of coordinating and retrieval project document significantly facilitate knowledge transfer in IT outsourcing at public agencies

H3d: The project routines of allocating and storing project document significantly facilitate knowledge transfer in IT outsourcing at public agencies

5. Final Remarks

This conceptual paper proposed an integrative preliminary framework that links four groups of key domains namely; client-related characteristics, vendor-related characteristics, Transactive Memory System context and project nature factors while discussing the theories and models behind the proposed model. This conceptual model is still based on literature study. Therefore, it needs further research to empirically validate the model. We believed that the application of the framework may provide useful insights into ITO specifically for Malaysia e-government initiatives.

References

[1] Hacket, B., Beyond Knowledge Management: New Ways to work and Lear, The Conference Board, Research Report, 1262-00-RR, 2000.

[2] MAMPU, Malaysia Public Agencies IT Outsourcing Guideline, 2006, URL: http://www.mampu.gov.my/pdf/Garis-Panduan-IT-outsource.pdf. [Cited 26 October 2010].

[3] Yusof, Z. and Ismail. M. B., Is There A Relationship Between Knowledge Sharing Practice And The Quality of Service Delivery? A Case Study In Three Governemnet Agencies In Malaysia, *Journal of Konwledge Managemeny, 10(1)*, 2009.

[4] King, R.C., Xia, W., Quick, J.C., and Sethi, V., Socialization and organizational outcomes of information technology professionals, *Journal of Career Development International, 10(1)*, pp. 26-51, 2005.

[5] Ko, D., Kirsch, L. and King, W., Antecedents of knowledge transfer from consultants to clients in enterprise system implementation, *MIS Quarterly, 29(1)*, pp. 59–85, 2005.

[6] Agrote, L., Ingram, P., Levine, J. M. and Moreland, R. L., Knowledge Transfer in Organizations: Learning from the Experience of Others, *Organizational Behavior and Human Decision Processes, 82(1)*, pp. 1-8, 2000.

[7] Schwartz, D.G., Integrating knowledge transfer and computer-mediated communication: categorizing barriers and possible responses, *Knowledge Management*, pp. 249-259, August 2007.

[8] Kumar J. A. and Ganesh, L. S., Research on knowledge transfer in organizations: A morphology, *Journal of Knowledge Management*, *13(4)*, pp. 161-174, 2009.

[9] Ambos, T.C. and Ambos, B., The impact of distance on knowledge transfer effectiveness in multinational corporations, *Journal of International Management*, *15(1)*, pp. 1-14, 2009.

[10] Szulanski, G., The process of knowledge transfer: a diachronic analysis of stickiness, *Organizational Behaviour and Human Decision Processes, 82(1)*, pp. 9-27, 2000.

[11] Mowery, D.C., Oxley, J.E. and Silverman, B.S., Strategic Alliances and Interfirm Knowledge Transfer, *Knowledge Creation Diffusion Utilization*, *17*, pp. 77-91, 1996.

[12] Al Ghassani, A.M., Improving the Structural Design Process: a Knowledge Management Approach, PhD thesis, Loughborough University, 2003.

[13] Connell J. and Voola, R., Strategic alliances & knowledge sharing: Synergies or silos?, *Journal of Knowledge Management , 11(3)*, pp. 52-66, 2007.

[14] Lee, J. N., The impact of knowledge sharing, organizational capability and partnership quality on IS outsourcing success, *Information and Management, 38*, pp. 323-335, 2001.

[15] Syed-Ikhsan, S.O.S.B. and Rowland, F., Benchmarking knowledge management in a public organisation in Malaysia, *Benchmarking: An International Journal*, *11*, 2004, pp. 238-266.

[16] Wong, K. W., An exploratory study on knowledge management adoption in the Malaysian industry, *International Journal of Business Information System, 3(3)*, pp. 272-283, 2008.

[17] Tat, L.W. and Hase, S., Knowledge Management in The Malaysian Aerospace Industry, *Journal of Knowledge Management, 11(1)*, pp. 143-151, 2007.

[18] Ali H.M. and Ahmad, N. H., Knowledge Management in Malaysian Banks: A New Paradigm, *Journal of Knowledge Management Practice, 7(3)*, 2006.

[19] Wei, C.C., Choy C.S. and Yew, W.K., Is the Malaysian telecommunication industry ready for knowledge management implementation?, *Journal of Knowledge Management*, *13(1)*, pp. 69 – 87, 2009.

[20] Sharimllah Devi, Chong, R. S.C. and Lin, B., Organizational culture and KM processes from the perspective of institution of higher learning, *International Journal of Management in Education, 1(1/2)*, pp. 57-79, 2007.

[21] Mohamed, A., Arshad, N. H. andAbdullah, N. A., Influencing factors of knowledge transfer in IT outsourcing, *Proceedings of the 10th WSEAS international conference on Mathematics and computers in business and economics*, pp. 165-170, 2009.

[22] Yang, B., LI, Q. and Zuo, M., Analysis on E-Government Outsourcing and its Model, *IFIP International Refedarion for Information Processing*, IEEE, pp. 1191-1195, 2008.

[23] Moon, J., Jung, G., Chung, M. and Choe, Y. C., IT outsourcing for E-government: Lessons from IT outsourcing projects initiated by agricultural organizations of the Korean government, *40th Annual Hawaii International Conference on System Sciences (HICSS'07)*, pp. 104a, 2007.

[24] Nakatsu, R.T. and Iacovou, C. L., A comparative study of important risk factors involved in offshore and domestic outsourcing of software development projects: A two-panel Delphi study, *Information and Management, 46*, pp. 57-68, 2009.

[25] Giannakis, M., Facilitating learning and knowledge transfer through supplier development, *Supply Chain Management: An International Journal, 13(1)*, pp. 62-72, 2004.

[26] Chen, Y. and Gant, J., Transforming local e-government services: the use of application service providers, *Government Information Quarterly, 18,* pp. 343-353, 2001.

[27] Benedikt, T. and Frank. M., Why risk management matters in it outsourcing – a systematic literature review and elements of a research agenda, *17th European Conference on Information Systems*, 1-13, 2009.

[28] Barney, J., Firm resources and sustained competitive advantage, *Journal of Management, 17(1)*, pp. 99-120, 1991.

[29] Wernerfelt, B., A resource-based view of the firm, *Strategic Management Journal, 5*, pp. 171-80, 1984.

[30] Li M. and Li, D., A Survey and Analysis of the Literature on Information Systems Outsourcing, *Pacific Asia Conference on Information Systems*, 2009.

[31] Combs, J. and Crook, T., Sources and consequences of bargaining power in supply chains, *Journal of Operations Management, 25*, pp. 546-55.

[32] Blumenberg, S., Wagner, H. and Beimborn, D., Knowledge transfer processes in IT outsourcing relationships and their impact on shared knowledge and outsourcing performance, *International Journal of Information Management, 29, pp*. 342-352, 2009.

[33] Joshi, K.D., Sarker, S. and Sarker, S., Knowledge transfer within information systems development teams: examining the role of knowledge source attributes, *Decision Support Systems, 43*, pp.322-334, 2007.

[34] Schramm, W., The Process and Effect of Mass Communication, Urbana: University of Illinois Press, 1954.

[35] Jacobson, C.M., Knowledge sharing between individuals, in Schwartz, D.G. (Eds), Encyclopedia of Knowledge Management, Idea Group Reference, Hershey, PA, pp.507-14, 2006.

[36] Hinsz, V.B., Tindale R.S. and Vollrath, D.A., The emerging conceptualization of groups as information processors, *Psychological Bulletin, 121*, pp. 43-64, 1997.

[37] Nonaka, I., A Dynamic Theory of Organizational Knowledge Creation, *Organization science, 5(1)*, pp. 14-37, 1994.

[38] Schramm, W., Mass Communication: a Book of Readings, Urbana: University of Illinois Press, 1960.

[39] Gibson, C.B., From knowledge accumulation to accommodation: cycles of collective cognition in work groups, *Journal of Organizational Behavior, 22*, pp. 121-34, 2001.

[40] Nonaka, I. and Takeuchi. H. The Knowledge-creating Company. Oxford University Press: New York, 1995.

[41] Sherif, K. and Xing, B., Adaptive processes for knowledge creation in complex systems: The case of a global IT consulting firm, *Information and Management, 43(4)*, pp. 530-540, 2006.

[42] Gourlay, S., Conceptualizing knowledge creation: A critique of Nonaka's theory. *Journal of Management Studies*, 43(7), 1415–1436, 2006.

[43] Curseu, P.P., Schalk, R. and Wessel, I., How do virtual teams process information? A literature review and implications for management, *Journal of Managerial Psychology, 23(6)*, pp. 628-652, 2007.

[44] Zhao, Z.J. and Anand, J., A multilevel perspective on knowledge transfer: evidence from the chinese automotive industry, *Strategic Management Journal, 30*, pp. 959-983, 2009.

[45] Easterby-smith, M., Lyles, M.A and Tsang, E.W.K., Inter-Organizational Knowledge Transfer: Current Themes and Future Prospects, *Journal of Management Studies*, *45*, pp. 677-690, 2008.

[46] Wilkesmann, U., Fischer, H. and Wilkesmann, M., Cultural characteristics of knowledge transfer. *Journal of Knowledge Management*, *13(6),* 464-477, 2009.

[47] Gregory, R., Beck, R. and Prifling, M., Breaching the Knowledge Transfer Blockade in IT Offshore Outsourcing Projects – A Case from the Financial Services Industry, *Proceedings of the 42nd Hawaii International Conference on System Sciences*, IEEE , pp. 1-10, 2009.

[48] Ochoa, S.F., Herskovic, V. and Pineda, E., A transformational model for Organizational Memory Systems management with privacy concerns, *Information Sciences, 179(15)*, pp. 2643-2655, 2009.

[49] Telvin Goh C.H. and Hooper, V., Knowledge and information sharing in a closed information environment, *Journal of Knowledge Management*, 13, pp. 21-34, 2009.

[50] Oshri, I., Fenema, P.C.V. and Kotlarsky, J., Knowledge Transfer in Globally Distributed Teams: The Role of Transactive Memory, *Information Systems Journal*, *18,* pp. 593-616, 2008.

[51] Jackson, P. and Klobas, J., Transactive memory systems in organizations: Implications for knowledge directories, *Decision Support Systems*, *44*, pp. 2409-424, 2008.

[52] Kotlarsky, J., Van Den Hoopp, B. and Huysman, M., The role of a transactive memory system in bridging knowledge boundaries, *Proceedings of the Organisational Learning, Knowledge and Capabilities (OLKC). Amsterdam*, The Netherlands; 2009.

[53] Narteh, B., Knowledge transfer in developed-developing country inter rm collaborations: a conceptual framework, *Journal of Knowledge Management*, *12(1)*, pp.78-91, 2008.

[54] Inkpen, A.C., Learning through joint ventures: a framework of knowledge acquisition, *Journal of Management Studies, 37,* 2000.

[55] Stein E.W. and Zwass, V., Actualizing organizational memory with information systems, *Information Systems Research*, *6(2)*, pp. 85-117, 1995.

[56] Nevo D. and Wand, Y., Organizational memory information systems: a transactive memory approach, *Decision Support Systems, 39*, pp. 549 – 562, 2005.

[57] Wegner, D.M., Erber, R. and Raymond, P., Transactive memory in close relationships, *Journal of Personality Social Psychology.*, *61*, pp. 923-929,1991.

[58] Anand, V., Manz, C.C. and Glick, W.H., An organizational memory approach to information management, *Academy of Management Review*, pp. 90–111, 1998.

[59] Keith, M. Demirkan, H. and Goul, M., Understanding Coordination in IT Project-Based Environments: An Examination of Team Cognition and Virtual Team Efficacy, *Proceedings of the 42nd Hawaii International Conference on System Sciences*, 2009.

[60] Lewis, K., Measuring transactive memory systems In the field: Scale development validation, *Journal of Applied Psychology*, 88, pp. 587-604, 2003.

[61] Kotlarsky, J. and Oshri, I., Social ties, knowledge sharing and successful collaboration in globally distributed system development projects, *European Journal of Information Systems*, *14*, pp. 37-48, 2005.

[62] Roberts, J., Analysis, T. and Management, S., From know-how to show-how? Questioning the role of information and communicat, *Technology Analysis and Strategic Management*, *12*, pp. 429-443, 2000.

[63] Aurum, A., Daneshgar, F. and Ward, J., Investigating Knowledge Management practices in software development organisations – An Australian experience, *Information and Software Technology*, *50*, pp. 511-533, 2008.

[64] Parent, R., Roy, M. and St-jacques, D., A systems-based dynamic knowledge transfer capacity model, *Journal of Knowledge Management*, *11*, pp. 81-93, 2007.

[65] Mu, J. Tang, F. and Maclachlan, D.L., Absorptive and disseminative capacity: Knowledge transfer in intra-organization networks, *Expert Systems With Applications*, *37*, pp. 31-38, 2010.

[66] Xu, Q. and Ma, Q., Determinants of ERP Impllementation Knowledge Transfer, *Information and Management*, *45*, pp. 528-539, 2008.

[67] Chen J. and Mcqueen, R.J., Knowledge transfer processes for different experience levels of knowledge recipients at an offshore technical support center, *Information Technology and People*, *23*, pp. 54-79, 2010.

[68] Pérez-Nordtvedt, L., Kedia, B.L., Datta, D.K. and Rasheed, A.A., Effectiveness and Effciency of Cross-Border Knowledge Transfer: An Empirical Examination, *Journal of Management Studies*, *45*, pp. 714-744, 2008.

[69] Cantu, L.Z. Criado, J.R. and Criado, A.R., Generation and transfer of knowledge in IT-related SMEs, *Journal of Knowledge Management*, *13*, pp. 243-256, 2009.

[70] Ajmal, K.U. and Koskinen, M. M., Knowledge Transfer in Project-Based Organizations: An Organizational Culture Perspective, *Project Management Journal*, *39*, pp. 7-15, 2008.

[71] Rhodes, J., Hung, R., Lok, P., Lien, B.Y.- hui and Wu, C.-min, Factors influencing organizational knowledge transfer: implication for corporate performance, *Journal of Knowledge Management*, *12*, pp. 84-100, 2008.

[72] Dhanaraj, C. Lyles, M.A. Steensma, H.K. and Tihanyi, L., Managing tacit and explicit knowledge transfer in IJVs: the role of relational embeddedness and the impact on performance, *Journal of International Business Studies*, *35*, pp. 428-42, 2004.

[73] Uzzi B. and Lancaster, R., The role of relationship in inter-firm knowledge transfer and learning the case of corporate debt markets, *Management Science*, *49(4)*, pp. 383-399, 2003.

TAXONOMY AND WIKIPEDIA TERMINOLOGY IDENTIFICATION AND EXTRACTION FROM AN UNSTRUCTURED TEXT

Hui-Ngo GOH and Ching-Chieh KIU

Faculty of Information Technology, Multimedia University, Jalan Cyberjaya, 63100 Cyberjaya, Selangor, Malaysia.

hngoh@mmu.edu.my

Abstract. This paper presents the use of a context-based methodology for term identification and extraction from an unstructured text document. The methodology used taxonomy and Wikipedia to support automatic term identification and extraction with an assumption of candidate terms for a topic is often affiliated with its topic-specific keywords. A hierarchical relationship of super-topics and sub-topics is defined by a taxonomy, meanwhile, Wikipedia is used to provide context and background knowledge for topics that defined in the taxonomy to guide the term identification and extraction. The experimental results have shown that context-based methodology for term identification and extraction is viable in defining topic concepts and its sub-concepts. The experimental results have also proven its viability to be applied in a small corpus / text size environment in supporting term extraction.

Keywords: *Term identification and extraction; Taxonomy; Wikipedia*

1. Introduction

Term identification and extraction emphasized in concept formation. It is a process of identifying significant terms that symbolize the content of the domain chosen. Therefore, the accuracy of extracted significant terms greatly impacts the usefulness of the intended practical application. Domain specific dictionary / thesaurus construction, document indexing for its categorical representation for easy retrieval, ontology construction are some of the typical examples.

To date, many research works have been carried to improve the outcome of term identification and extraction. They are categorized into five approaches which are linguistic-based, statistical-based, machine learning-based ontology-based and hybrid-based. Linguistic-based approach [1], [2], [3] uses dictionary, rules or patterns to extract desired terms; high workload is imposed on linguists as all rules and patterns have to be coded manually but it produces better quality outcome. Meanwhile, statistical approach [4, [5], [6], [7], [8], [9], [10], [11] uses solely mathematical computation to identify the co-occurrence of lexical(s) in one or more specified set of documents. Generally, it performs solely on the term counting without any semantic understanding. On the other hand, the advantage of machine learning approach [12], [13], [14] is that it is relatively easy to tune to new domains, provided that tagged training data is existed. [15], [16] have adopted ontology in extracting terms, where it can achieve higher performance results; however, it requires the availability of the domain specific ontology prior the extraction. Lastly, hybrid approach is the combination of various approaches in performing term identification and extraction. For example, C/NC value [17], a hybrid approach that use linguistic-based to form terms' patterns for identification and extraction before statistically counting its contextual information. However, all above mentioned term identification and extraction approaches are mainly constrained by four principal as discussed below.

a) *Domain Level*: Most research works required the selection of corpus resources. The selection of corpus is usually domain specific to the intended problem to be solved and a domain corpus might consist of various distinct topics. The above mentioned approaches are solely focused on one level (domain level) term identification and extraction, where it often accommodates to the nature of the investigated domain as a whole context. This might constraint its identification and extraction outcome/result when applying it to other domains or various topics as they might be different in context and background knowledge.

b) *Nature of the terms*: Research works done in term identification and extraction can be categorised into two areas, which are technical and non-technical. In the technical area, it implies the use of specialized knowledge of applied sciences such as for medicine and biology domain. Meanwhile, the non-technical area denotes the use of general knowledge such as for tourism and educational domain. Both areas exhibit the increasing usage of diversity in term of morphology and collocation. Despite to the distinct nature of terms between the technical area and the non-technical area, the technical areas often expose certain pattern in its terminological presentation, for instance, biological terms often contain prefixes and suffixes that give an indication of their class. On the other hand, the non-technical areas are always clueless to accommodate precisely the likelihood of potential terms.

c) *Text / corpus size*: Research works done in term identification and extraction often involve multi-documents with the aim to conciliate the relevancy of extracted terms to the domain investigated. Various statistical metrics are then used to validate the extracted terms relevancy to the domain chosen. Frequency-based counting and Term Frequency – Inverse Document Frequency (TF-IDF) are the two most commonly used metrics in validating true terms. In this case, corpus size does impact the term identification and extraction. However, it will be a problematic issue for domains which have less resources and small corpus size in term extraction.

d) *Type of text/ corpus*: All approaches mentioned above are mainly focus at monolingual term identification and extraction. It is one language of terms extraction. However, there will be a difficulty if a text contains more than one language. For instance, English and Malay words appear in the same text.

In this paper, we propose a context-based term identification and extraction using taxonomy and Wikipedia to overcome the above mentioned issues. A hierarchical relationship of super-topics and sub-topics is defined by a taxonomy, meanwhile, Wikipedia is used to provide context and background knowledge for topics that defined in the taxonomy to guide the term identification and extraction.

The paper is structured as follows: Section 2 describes the use of taxonomy and Wikipedia in term identification and extraction, system framework is presented in section 3 where each individual steps are discussed in detail, section 4 gives the experiments, evaluations and discussions on the proposed work, section 5 explains the limitation and possible future works and finally section 6 concludes this paper.

2. Background Technologies

2.1. Taxonomy

To date, various research works have been carried out on term extraction, however their works are mainly focus at domain level extraction. As defined, a domain might consist of various topics. For example, *Tourism* domain comprises of *culture and heritage, hotel*, *transportation* and *places to visit* as its topics.

In our work, taxonomy is proposed to be used as it provides a structure for various topics in a domain. It defines a hierarchical relationship of super-topics and sub-topics. Hence, the domain level term identification and extraction is performed by taking into the consideration of different topics might appear in the investigated domain in which topic related documents will be handled specifically according to its context and background during term identification and extraction.

- About Malaysia
    - Culture and Heritage
        - People
        - Games
        - Music
        - Dance
        - Textiles
    - Essentials
        - Transportation
        - Passport/Visa

Figure 1. Domain taxonomy on Malaysia tourism

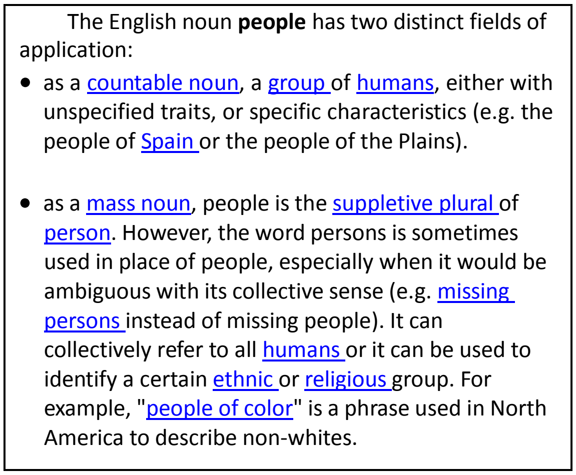
Figure 1 indicates a portion of taxonomy for Malaysia tourism as an example. As illustrated in the Figure 1, the Malaysia tourism domain/webpage consists of various topics; each topic is distinct in its contents. For example, the parent concept "*About Malaysia*" consists of "*Culture and Heritage*" and "*Essentials*" as the first level sub-topic in domain taxonomy; meanwhile "*People*", "*Games*", "*Music*", "*Dance*" and "*Textiles*" are represented as the second level sub-topics of "*Culture and Heritage*" whereas "*Transportation*" and "*Passport/Visa*" are represented as the second level sub-topic of "*Essentials*" in the domain taxonomy.

2.2. Wikipedia

Wikipedia is the world largest online encyclopedia lies in its size and coverage. It reaches approximately 3 million articles in English as dated on May 2010 since its establishment in year 2001. It covers a rich resource of general knowledge as well as in depth clarification of many specialized knowledge which might be potentially contribute in various aspects to knowledge extraction.

In our work, Wikipedia is used to provide context and background knowledge to topics in a domain taxonomy. An assumption is formed where candidate terms of a topic are often associated with its topic-specific keywords in providing a context and background knowledge.

Wikipedia article corresponding to the topics represented in domain taxonomy is elicited. Hyperlinks exist in each Wikipedia article symbolizes *descriptive words* (*dw*); it is the description of context and background knowledge to the investigated topic. Figure 2 shows the introduction part of the topic "*People*" described in Wikipedia. The underline words (hyperlinks) are example of its *dw*. Figure 3 is the list of extracted *dw* for topic "*people*" and it symbolized the topic "*people*" context and background knowledge.

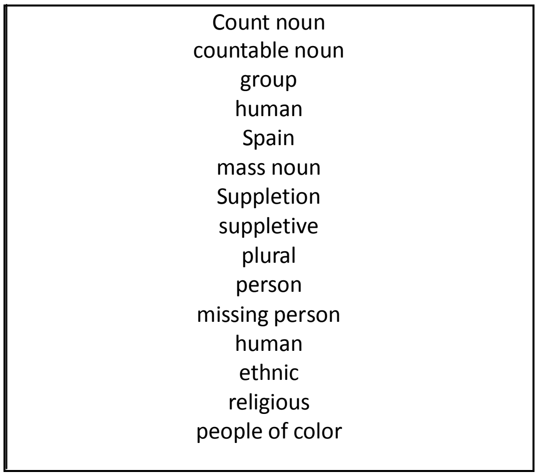


The English noun **people** has two distinct fields of application:

- as a countable noun, a group of humans, either with unspecified traits, or specific characteristics (e.g. the people of Spain or the people of the Plains).

- as a mass noun, people is the suppletive plural of person. However, the word persons is sometimes used in place of people, especially when it would be ambiguous with its collective sense (e.g. missing persons instead of missing people). It can collectively refer to all humans or it can be used to identify a certain ethnic or religious group. For example, "people of color" is a phrase used in North America to describe non-whites.

Figure 2. Wikipedia content on topic "People"



Count noun
countable noun
group
human
Spain
mass noun
Suppletion
suppletive
plural
person
missing person
human
ethnic
religious
people of color

Figure 3. Extracted *dw* on topic "People"

3. System Framework

The prototypical implementation of context-based term identification and extraction of our approach is illustrated in Figure 4 and it details algorithm is formulated as below:

Step 1 : Identify domain source
A domain is identified as input source for term identification and extraction. For the experimental purpose, tourism domain is chosen to test the proposed framework. Domain web pages are taken from Malaysia official tourism website[2] as the source documents for term identification and extraction. "*People*", "*Games*", "*Music*", "*Dance*" and "*Textiles*" are the selected topics with their corresponding web page in the selected domain web page.

Step 2 : Identify domain taxonomy
Given the identified domain, a related domain taxonomy is defined. In the experiment, a domain taxonomy corresponding to the domain is adopted from Malaysia official tourism website. Its hierarchical relationship of super-topics and sub-topics defined by the taxonomy as illustrated in Fig 1. is to provide structure for term identification and extraction from structured documents.

Step 3 : Retrieve related Wikipedia articles of the topics defined in the domain taxonomy
A freely available Wikipedia API[3] is used to provide automatic access to Wikipedia articles. For example, topic's name, "*People*" in the domain taxonomy is served as a keyword for retrieving a related Wikipedia article. The Wikipedia API implements OpenSearch protocol to retrieve Wikipedia article using the provided keyword. The related Wikipedia article(s) corresponding to the provided keyword is returned.

[2] http://www.tourism.gov.my
[3] http://en.Wikipedia.org/w/api.php

Step 4 : Generation of the list of descriptive words

The obtained Wikipedia article (webpage) is rendered into HTML format automatically using info.bliki.api.creator[4], a package in the Wikipedia API (Bliki engine)[5]. The objective of this step is to ease the generation of the list of descriptive words. A set of a word list is defined as follow:

$$DW = \{dw_1, dw_{2,}, \ldots dw_n\}$$

where *DW* denotes a list of descriptive word (*dw*) as described in Figure 3. All hyperlinks (descriptive words) will be extracted from the converted HTML document. The below HTML syntax is the sample hyperlink of a related descriptive word.

<a href="/wiki/Person" title="Person">person</a>

<a> …. </a> is the anchor name used to display information within a document, *href="/wiki/Person* denotes the URL of the descriptive word, and *title="Person"* indicates title/description of the descriptive word. *"person"* is a descriptive word as explained in Section Wikipedia. In this case, "*person*" is extracted.

Step 5 : Document cleaning

"*People*", "*Games*", "*Music*", "*Dance*" and "*Textiles*" are the topics available in the selected domain web pages as stated in step 1. Each topic webpage is retrieved from the domain web page and cleaned up using HTML Content Extractor[6] to eliminate non-text contents such as ads, banners, videos, audios, navigations links and menus. The cleaning task is performed automatically and does not require any user interaction during the cleaning process. At the end of the process, a pure text file of the topic is produced.

Step 6 : Term identification

Given an assumption that the candidate terms of a topic are often associated with its topic-specific keywords, each *dw* in the list of descriptive word is examined against each sentence in pure text file of the topic. Sentence which contains *dw* is extracted. Finally, a file with sentences where each sentence contains at least one *dw* is being generated.

Step 7 : Term extraction

This step consists of two processes which are tagging and stemming.

a. Tagging

Treetagger[7] (a multi-lingua tool for annotating text with part-of-speech and lemma information) is used to shallowly tag all the extracted sentences in the step 5 and to elicit terms which are tagged with "NP" (Noun Phrase). Generally, most candidate terms possess to be tagged with Noun Phrase. A list of terms is generated as an input for next process.

b. Stemming

Extracted list of terms might contain redundancy. For example, the word "*malays*" and "*malay*" are in fact referring the same item. Hence, Porter Stemmer[8] is used to reduce all terms into their stem, base or root form. In this case, after the stemming process, the base form of "*malays*" and "*malay*" is "*malai*". The terms with the same base form will be considered as one term.

[4] http://matheclipse.org/doc/bliki/index.html

[5] http://www.matheclipse.org/en/Java_Wikipedia_API

[6] http://senews.sourceforge.net/KCE_README.html

[7] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[8] http://drupal.org/project/porterstemmer

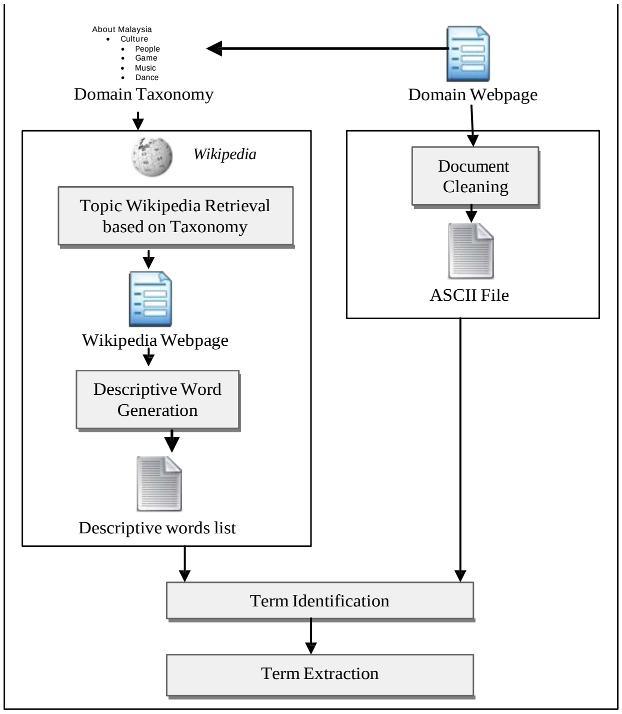


Figure 4. System framework

4. Experiment

4.1. Dataset

Malaysia official tourism website was used to test our approach. The available domain taxonomy on the designated Malaysia tourism website is adopted to guide the term identification and extraction.

Five distinguished topics in the domain taxonomy, "*People*", "*Games*", "*Music*", "*Dance*" and "*Textiles*" as illustrated in Figure 1 are chosen for testing our proposed approach. The content of each topic is context sensitive to its title. Thus, it is significantly novel to experiment the taxonomy utilization in term identification and extraction.

Table 1 displayed the word count for each topic in a pure text file.

Table 1. Topic word count

Topic	word count
People	937
Games	671
Music	282
Dance	873
Textiles	196

4.2. Evaluation metrics

The most challenging activity in term extraction lies in its evaluation method as there is no formal way to evaluate terms. Same resource corpus might produce different terms, all are depends on their usage in the developed application. Hence, it is difficult to obtain a suitable "gold standard" that can be used to evaluate extracted terms.

Having all the known evaluation difficulties in mind, we manually evaluated the result with the help of human expert. The term identification and extraction were evaluated using two metrics: Precision and Recall as shown in Eq. 1 and Eq. 2.

$$\text{Precision} = \frac{\text{No of correctly extracted term}}{\text{Total terms extracted by the system in investigated topic}} \quad (1)$$

$$\text{Recall} = \frac{\text{No of correctly extracted term}}{\text{Total terms exist in the investigated topic}} \quad (2)$$

4.3. Result and discussion

Table 2. Recall and precision of term extraction

Topic	Recall	Precision
People	100	60
Games	62.5	38
Music	62.5	41.7
Dance	100	47.5
Textiles	100	44

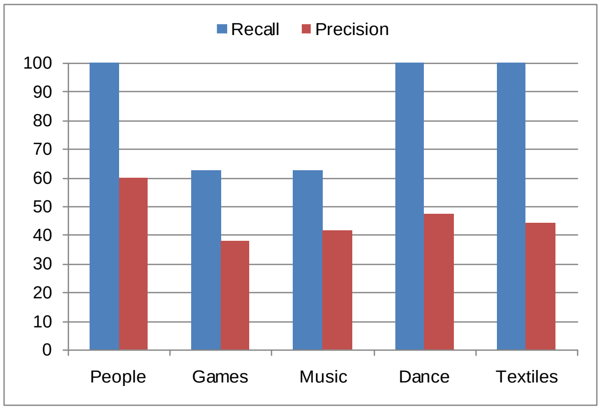


Figure 5. Recall and precision of term extraction

The results of term extraction are presented in Table 2 and Figure 5, they showed that the domain taxonomy and Wikipedia have contributed an impact in our work. The use of *dw* in topic web page hints the exact location of the topic candidate terms. For instance, "*humans*", "*person*" and "*ethnic*" are the most referred *dw* against topic "*People*". Whenever the descriptive word occurs in a sentence, the topic candidate term was existed before or after it. Out of 60% of the extracted sentences in the topic "*People*", it contains 100% of terms in it. This has proven that context-based of using domain taxonomy and Wikipedia worked well in handling term identification and extraction by taking into the consideration of different topics might appear in the investigated domain.

This experiment has also discovered that the page size of topic has not contributed to its performance metrics. The largest size of topic webpage, "*People*" with 937 word count gives 100% recall and 60% precision whereas the smallest topic webpage, "*Textiles*" with only 196 word count gives a recall of 100% and precision of 44%. This gives an indication that our approach can be applied in any domain regardless of the text / corpus size.

4.4. Comparison evaluation result and discussion on term extraction

As indicated in Table 3 and Figure 6, we performed a comparison study between our approach with Yahoo! Search Keyword Terms Extraction[9] and Termine[10]. Yahoo Search Keyword Terms Extraction employs Yahoo! Search API, whereas Termine adapts C-value, a statistical approach to perform term extraction. Both tools were chosen as they share the same construction characteristics with our approach in term of domain portability and scalability, and they are also an automatic term identification and extraction tool without the need of any training corpus.

In our approach, the extracted sentences generated from the term identification process are tagged using Treetagger. The terms which are tagged with "NP" aka Noun Phrase is extracted.

As can be seen from the Table 3 and Figure 6, our approach is outperformed than Yahoo Search Keyword Terms and Termine in term of precision and recall. The methodology of the context-based term identification and extraction using taxonomy, and Wikipedia performed domain level term identification and extraction by taking into the consideration of different topics might appear in the investigated domain compared to the both tools which have performed the term extraction on the domain level regardless of the context and background knowledge of its content.

Figure 7, 8 and 9 are the results of the terms identification and extraction using Yahoo! Search Keyword Terms Extraction, Termine and our approach on the topic "People". As shown by the results, only the Termine contains value for each term as they are extracted based on statistical approach and the extracted terms are listed according to the calculated value. Compared to the Yahoo Search keyword Terms and Termine, our approach is able to extract significant language (English and Malay) terms which are in a proper collocation.

Table 3. Recall and precision of term extraction tools

	Yahoo! Keyword Phrase Term Extraction		Termine		Our Approach	
Taxonomy concept	**Recall**	**Precision**	**Recall**	**Precision**	**Recall**	**Precision**
People	8.3	5	16.7	2.5	100	40
Games	0	0	62.5	7.6	62.5	31.3
Music	12.5	5	37.5	6.5	37.5	19
Dance	10.5	10	47.3	7.8	100	46.3
Textiles	0	0	50	8.7	75	60

[9] http://tools.seobook.com/yahoo-keywords
[10] http://www.nactem.ac.uk/software/termine

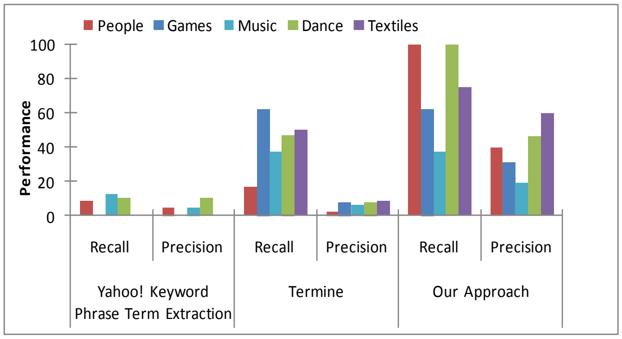


Figure 6. Performance among term extraction tools

```
<ResultSet xsi:schemaLocation="urn:yahoo:cate
http://api.search.yahoo.com/ContentAnalysisService/V1/TermExtractionResponse.xsd">
<Result>malay traditions</Result>
<Result>theravada buddhism</Result>
<Result>keen business sense</Result>
<Result>state of johor</Result>
<Result>colourful heritage</Result>
<Result>conversion to islam</Result>
<Result>malay language</Result>
<Result>spice route</Result>
<Result>chinese immigrants</Result>
<Result>indigenous ethnic groups</Result>
<Result>malaysian culture</Result>
<Result>tourism malaysia</Result>
<Result>malays</Result>
<Result>hokkien</Result>
<Result>1400s</Result>
<Result>spea</Result>
<Result>ethnic groups in malaysia</Result>
<Result>mannerisms</Result>
<Result>sarawak</Result>
<Result>chinese language</Result>
</ResultSet>
```

Figure 7. Yahoo! Search Keyword Terms Extraction on topic “People”

```
Rank  Term                                                      Score
1     ethnic group                                              11.5
2     indigenous ethnic group                                   3.169925
3     indigenous ethnic groups orang asli orang asli            2.807355
4     tamil-speaking south indian immigrant                     2
4     annual tamu besar festival                                2
4     orang ulu                                                 2
4     malaysian culture                                         2
4     capital city kuala lumpur                                 2
4     kadazan dusun                                             2
```

4	mountainous region	2
4	upriver tribe	2
12	british colonial rule	1.584962
12	keen business sense	1.584962
12	nomadic sea-faring people	1.584962
12	malaysian chinese form	1.584962
12	flat valley delta	1.584962
12	rich art heritage	1.584962
12	intriguing diversity malay	1.584962
12	indian caste system	1.584962
12	kadazan dusuns form	1.584962
12	paddy field farming	1.584962
12	fearsome warrior race	1.584962
12	islamic coastal population	1.584962
12	impressive equestrian skill	1.584962
12	main ethnic group	1.584962
12	northern inland region	1.584962
12	international spice route	1.584962
12	main tribal group	1.584962
12	traditional community home	1.584962
12	orang ulu tribe	1.584962

Figure 8. Termine on topic "People"

Malaysia
Islam
CHINESE
Hokkien
Penang
Kuala Lumpur
Johor
MALAY
INDIAN
ETHNIC GROUPS
Orang Asli
Peninsular Malaysia
Negrito
Senoi
Iban
Bidayuh
Orang Ulu
Sarawak
Dayak
Sabah
Kadazan
Dusun
Bajau
Murut
Borneo

Figure 9. Our approach on topic "People"

5. Limitation and Future Enhancement

The limitation our proposed framework is the over generation of topic's *DW*. For example, as can be seen in Figure 3, the *DW* of "*Count noun*", "*countable noun*", "*Spain*", "*mass noun*", "*Suppletion*", "*suppletive*" and "*plural*" are not significant to the context and background knowledge of the topic "*People*". They are merely appeared in the *DW* due to its hyperlink listed in the topic web page.

In the future work, we intend to look into synonym of the each *dw* and semantic interpretation of Wikipedia content in improving the performance of term identification and extraction.

6. Conclusion

The work proposed in this paper is meant to provide a better way for term identification and extraction by taking into consideration of different topics might occur in a domain corpus for supporting ontology construction. The multi-topics are represented in taxonomy as a multi-level tree representation and the Wikipedia is used to provide multi topics' context and background knowledge. The hypothesis proven is that carefully taking care of the need of each domain topics can improve the performance metrics.

7. References

[1] Bajwa. I. S., and Siddique. M. I., and Choudhary. M. A.:, "Automatic Domain Specific Terminology Extraction using a Decision Support System", In the Proceedings of 4th IEEE - International Conference on Information and Communication Technology-ICICT, pp. 651-659, Cairo, Egypt, 2006.

[2] Wermter. J., and Hahn, U., "Finding New Terminology in Very Large Corpora", Proceedings of the 3rd international conference on Knowledge capture Banff, pp. 137-144, Alberta, Canada, 2005.

[3] Mukherjea. S., and Subramaniam. L. V., and Chanda. G., and Sankararaman. S., and Kothari. R., and Batra. V., and Bhardwaj. D., and Srivastava. B., "Enhancing a Biomedical Information Extraction System with Dictionary Mining and Context Disambiguation", IBM Journal of Research and Development, Volume 48 , Issue 5/6, pp. 693 – 701, 2004.

[4] Chang. J.S., "Domain Specific Word Extraction from Hierarchical Web Documents: a First Step Toward Building Lexicon Trees from Web Corpora", Proceedings of the Fourth SIGHAN Workshop on Chinese Language Learning, Korea, pp.64-71, October 2005.

[5] Chen. Y. R., "The Research on Automatic Chinese Term Extraction Integrated with Unithood and Domain Feature", Master Thesis in Beijing, Peking University 2005.

[6] Kurz. D.F., and Xu. Y., "Text Mining for the Extraction of Domain Relevant Terms and Terms Collocations", Proceedings of the International Workshop on Computational Approaches to Collocations, Vienna, Austria, July 2002.

[7] Church. K. W., and Gale. W. A., "Concordances for Parallel Text", In Proceedings of the Seventh Annual Conference of the UW Center for the New OED and Text Research, Association for Computational Linguistics, Oxford, UK, pp. 40–62, 1991.

[8] Streiter. O., and Zielinski. D., and Ties. I., and Voltmer, L., "Term Extraction for Ladin: an Example Approach", TALN: Traitement Automatique des Langues Naturelles, VVF-Batz-sur-Mer(44), France, 2003.

[9] Dunning. T., "Accurate Methods for the Statistics of Surprise and Coincidence", Computational Linguistics, Vol. 19, no. 1, pp. 61–74, 1993.

[10] He. T.T., and Zhang X.P., and Ye X.H., “An Approach to Automatically Constructing Domain Ontology”, PACLIC 2006, Wuhan, China, pp. 150-157, 1-3 November, 2006.

[11] Alexander G., and Grigori S., and Eduardo LV., and Liliana C.H., “Automatic Term Extraction using Log-likelihood based Comparison with General Reference Corpus”, LNCS 6177, pp. 248-255, 2010.

[12] Eriksson. G., and Franzén. K., and Olsson. F., and Asker. L., and Lidén, P., “Exploiting Syntax when Detecting Protein Names in Text”, EFMI Workshop on Natural Language Processing in Biomedical Applications, Nicosia, Cyprus, March 2002.

[13] Zhang. Q.L., and Lu. Q., and Sui. Z.F., “Measuring Termhood in Automatic Terminology Extraction”, International Conference on Natural Language Processing and Knowledge Engineering, Beijing, China, pp.328 – 335, 2007.

[14] Zhou. G. D., and Su. J., “Named Entity Recognition using an HMM-based Chunk Tagger”, Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, pp. 473-480, 2002.

[15] Zhang. W., and Yoshida. T., and Tang, X.J., “Using ontology to improve precision of terminology extraction from documents”, Expert Systems with Applications, Vo. 36, Issue 5, pp. 9333-9339, 2009.

[16] Zhou. X.H., and Han. H., and Chankai. I., and Prestrud. A., and Brooks. A., “Approaches to Text Mining for Clinical Medical Rrecords”, Proceedings of the 2006 ACM symposium on Applied computing, Dijon, France, pp.235 – 239, 2006.

[17] Ananiadou. S., and Nenadic. G., “Automatic Terminology Management in Biomedicine”, Text Mining for Biology and Biomedicine, S. Ananiadou and J. McNaught (eds), Artech House, London, Ch.4, pp. 67-98, 2006.

EXPERIMENT ON CROSS-LINGUAL INFORMATION RETRIEVAL USING MALAY-ENGLISH BILINGUAL DICTIONARY

Nurul Amelina Nasharuddin, Muhamad Taufik Abdullah, Rabiah Abdul Kadir,
Azreen Azman and Nurjannaton Hidayah Rais
Department of Multimedia
Universiti Putra Malaysia
UPM Serdang, Selangor, Malaysia
E-mail:{amelina, taufik, rabiah, azreen}@fsktm.upm.edu.my, jannahhidayah@gmail.com

Abstract. The simplest way to search for the information is to look at every item in a collection of information or in database but the collection not necessarily be in one language only as information does not limited to language. When the need to translate the languages being used arises, this is where the techniques and methods that were being developed for the cross-lingual retrieval system will take place. This article reviews some recent researches and issues focusing on topics in cross-lingual information. An experiment has also being conducted in order to evaluate the effectiveness of using the bilingual dictionary in Malay-English cross-lingual retrieval system.

Keywords: Cross-lingual information retrieval, dictionary-based translation, query translation, bilingual dictionary, Malay language

1. Introduction

Information retrieval (IR) can be defined in so many ways. Asking the current time is one of the forms of information retrieval. In 2008, Manning, Raghavan and Schutze [1] define IR as "the act of finding material usually documents of an unstructured form (usually text) that satisfies an information need within large collections (usually stored in computers). These tasks not restricted to only documents in one language but also in other languages. But the classical IR regards the documents in foreign language as the unwanted "noise" [2]. These needs introduce new area of IR which takes into account all the documents received regardless of the languages being used; the cross-lingual and multi-lingual IR.

In classical IR search engines, both the query and the retrieved documents are in the same language. The retrieved documents in the cross-lingual IR system can be in a language different from the query language used by a user. But now rather than focusing only on two languages, IR system also can retrieved documents in multiple languages. Of course there are also lots of problems arises with these enhanced systems.

Cross-lingual IR has become more important in recent years. The basic idea behind the cross-lingual IR is to retrieve documents in a language (or called as the target language) different from the query language (or the source language) used by the user to develop the query. This may be desirable even when the user is not a speaker of the language used in the retrieved documents. Once the user knows that the information about their needs exist and is relevant, the retrieved documents can be translated by a human translator for future use of the user. Translations can be performed on the query, the document or in both document and query [3]. Query translation involves translating the query to the target language while document translation will translate whole document into source language.

The classical IR which is in mono-lingual is obviously proven successful since at present, searching has been the most used tool in the Web. However, when it comes to cross-lingual IR the situation is quite different. Lilleng and Tomassen [4] mention that there are few satisfactory quality cross-lingual IR systems available for the Web, but cross-lingual approaches for restricted domain (for example the medicine and geo-informatics) have shown to be more rewarding.

The structure of this article is as follows. Section 2 discusses current researches in cross-lingual IR which include the research using an improved dictionary, ontologies, interaction with user and expansion of the queries. Section 3 will describe a current research that focused on using the bilingual dictionary in translating user query for Malay and English language. Result of the experiment and the discussion will be discussed in Section 4 and this article will be summarized and a discussion on the future of cross-lingual IR will be described in Section 5.

2. Cross-lingual IR Current Researches

2.1. Improving Dictionary-based Cross-lingual IR

Using the dictionary based translation is a traditional approach in cross-lingual IR systems but significant performance degradation is observed when queries contain words or phrases that do not appear in the dictionary. This is called the Out-of-Vocabulary (OOV) problems [5]. This is to be expected even in the best of dictionaries. Input queries by user usually short and even the query expansion cannot help to recover the missing words because of information lacking. Generally, OOV terms are proper names or newly created words. For example, a user wants to search the information about the Influenza A (H1N1) disease in Malaysia by entering “H1N1 Malaysia” as the query. The H1N1 is a newly created term and may not be included in a dictionary which was published only a few years ago. If the term H1N1 is omitted from the query translation, it is most likely that the user will not get any relevant documents at all. OOV terms include compound words, proper nouns and technical terms [6].

In many documents technical terms and proper names are important text elements. Dictionaries only include the most commonly used proper nouns and technical terms used such as major cities and countries. Their translation is crucial for a good cross-language IR system. A common method used to handle un-translatable keywords is to include the un-translated in the target language query. If this word does not exist in the target language, the query will be less likely to retrieve relevant documents.

Named entities (NEs) are essential components of texts, especially news texts [7]. NEs extraction and translation are vital in the field of NLP for research on machine translation, cross-language information retrieval, bilingual lexicon construction, and so on. There are three types of NEs [8]; entity names (organizations, persons and locations), temporal expressions (dates and times), and number expressions (monetary values and percentages).

Organizations, person and location named entities are difficult to handle with a fixed set of rules, since new entity names are constantly being created and hence the growing need to investigate techniques for NEs extraction and translation. Bilingual dictionaries often have few entries for NEs [9]. But, when NEs are wrongly segmented as ordinary words and translated with a bilingual dictionary, the results are often poor.

Wikipedia [10] has becoming an important resource in the cross-lingual IR recently. Many researchers have conducted studies and experiments using the free online encyclopaedia. Lin, Wang, Yeh, Tsai and Tsai [9] have developed a Japanese-Chinese IR system based on the query translation approach. The system employs a more conventional Japanese-Chinese bilingual dictionary and Wikipedia for translating query terms. They studied the effects of using Wikipedia and proposed that Wikipedia can be used as a good NEs bilingual dictionary. By exploiting the nature of Japanese writing system, the query terms are processed differently based on the forms they are written in. To cope with term disambiguation, they have adopted an iterative disambiguating method based on the PageRank algorithm. The method proved to be effective and outperforms the previous Japanese-Chinese system’s tests.

A recent Wikipedia-based study by Nguyen, Overwijk, Hauff, Trieschnigg, Hiemstra, and de Jong [11] showed that query translations for cross-lingual IR can be performed using only Wikipedia. An advantage of using Wikipedia is that it allows translating phrases and proper nouns well. It is also very scalable since it is easy to use the most up to date version of Wikipedia which makes it able to handle actual terms. The approach is that the queries are mapped to Wikipedia concepts and the corresponding translations of these concepts in the target language are used to create the final query. The system build by the authors called WikiTranslate is evaluated by searching with topics in Dutch, French and Spanish in an English data collection.

The system which achieved a performance of 67% compared to the monolingual baseline can be valuable alternative to current translation resources and that the unique structure of Wikipedia (for example the text and internal links) can be very useful in CLIR. The use of Wikipedia might also be suitable for Interactive CLIR, where user feedback is used to translate the query, since Wikipedia concepts are very understandable for people.

2.2. Translation Resources beyond Human-constructed Dictionaries

Query suggestion aims to suggest relevant queries for a given query, which help users to specify their information needs better [12]. It is closely related to query expansion but the query suggestion aims to suggest full queries that have been formulated by users in another language. Gao et al [12] proposed query suggestion by mining relevant queries in different languages from up-to-date query logs as it is expected that for most user queries, we can find common formulations on these topics in the query log in the target language.

Therefore, cross-lingual query suggestion also plays a role of adapting the original query formulation to the common formulations of similar topics in the target language. Used as a query translation system, the proposed method demonstrates higher effectiveness than traditional query translation methods using either bilingual dictionary or machine translation tools.

2.3. Ontologies

Pourmahmoud and Shamsfard [13] carried out a research to retrieve English documents relevant to Persian queries using bilingual ontologies to annotate the documents and queries. A bilingual ontology consists of an ontology and a bilingual dictionary. Ontology is a formal, explicit specification of a shared conceptualization. It contains a set of distinct and identified concepts related by a set of relations [14]. They use the ontology to expand the query with related terms in pre and post translation expansion and the combined approach significantly improves cross-lingual performance. An advantage of this approach is the adaptability to several languages, which can be done by adding other dictionaries and thesauruses.

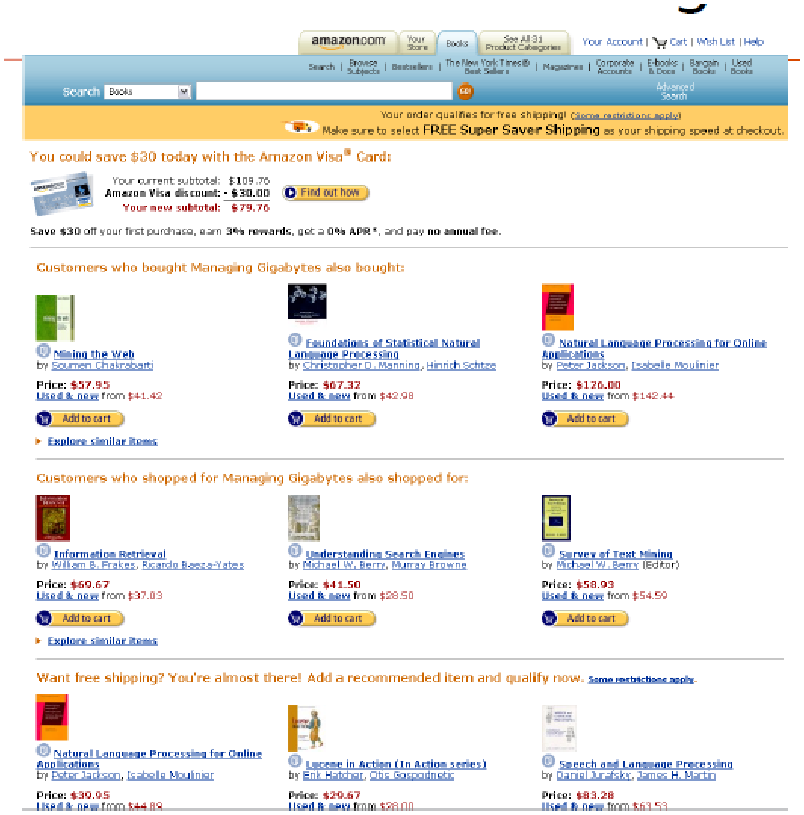
2.4. Interaction with the User

One of the criticisms frequently made of research in the CLIR domain is that too much attention has been given to questions of retrieval functionality and effectiveness with little regard to the real needs of the end user [15]. Interactive query expansion (IQE) [16] is a potentially useful technique to help searchers formulate improved query statements, and ultimately retrieve better search results. However, IQE is seldom used in operational settings because IQE is generally not integrated into searchers' established information-seeking behaviours (e.g., examining lists of documents), and it may not be offered at a time in the search when it is needed most (i.e., during the initial query formulation).

These challenges can be addressed by coupling IQE more closely with familiar search activities, rather than as a separate functionality that searchers must learn. White and Marchionini [17] introduced and evaluated a variant of IQE known as Real-Time Query Expansion (RTQE). As a searcher enters their query in a text box at the interface, RTQE provides a list of suggested additional query terms, in effect offering query expansion options while the query is formulated.

2.5. Query Expansion

Research done by Han and Chen [18] focusing on a new query expansion method which combined the ontology-based collaborative filtering and neural networks to increase the performance of expansion method. Collaborative filtering is a method that assesses similarities among the previous users then recommends the documents that are accessed by the users to the current users [19]. Example of collaborative filtering is in Fig. 1. But the limitation of the collaborative filtering is that it cannot find any connection between users and their similar users when little information was given by the user at the beginning of the search.


amazon.com
Your Store
Books
See All 31 Product Categories
Your Account | Cart | Wish List | Help
Search | Browse Subjects | Bestsellers | The New York Times® Best Sellers | Magazines | Corporate Accounts | E-books & Docs | Bargain Books | Used Books
Search Books
GO
Advanced Search

Your order qualifies for free shipping! (Some restrictions apply)
Make sure to select FREE Super Saver Shipping as your shipping speed at checkout.

You could save $30 today with the Amazon Visa® Card:

Your current subtotal: $109.76
Amazon Visa discount: - $30.00
Your new subtotal: $79.76

Save $30 off your first purchase, earn 3% rewards, get a 0% APR*, and pay no annual fee.

Customers who bought Managing Gigabytes also bought:

Mining the Web
by Soumen Chakrabarti
Price: $57.95
Used & new from $41.42

Foundations of Statistical Natural Language Processing
by Christopher D. Manning, Hinrich Schtze
Price: $67.32
Used & new from $42.98

Natural Language Processing for Online Applications
by Peter Jackson, Isabelle Moulinier
Price: $126.00
Used & new from $142.44

Customers who shopped for Managing Gigabytes also shopped for:

Information Retrieval
by William B. Frakes, Ricardo Baeza-Yates
Price: $69.67
Used & new from $37.03

Understanding Search Engines
by Michael W. Berry, Murray Browne
Price: $41.50
Used & new from $28.50

Survey of Text Mining
by Michael W. Berry (Editor)
Price: $58.93
Used & new from $54.59

Want free shipping? You're almost there! Add a recommended item and qualify now. Some restrictions apply.

Natural Language Processing for Online Applications
by Peter Jackson, Isabelle Moulinier
Price: $39.95
Used & new from $44.89

Lucene in Action (In Action series)
by Erik Hatcher, Otis Gospodnetic
Price: $29.67
Used & new from $28.00

Speech and Language Processing
by Daniel Jurafsky, James H. Martin
Price: $83.28
Used & new from $63.53

Fig. 1: An example of collaborative filtering from Amazon website

The two proposed query expansion approaches are thoroughly evaluated using two standard Text Retrieval Conference (TREC) Web collections and from the experimental results, there is a statistically significant improvement compared with the baselines. They concluded that the adaptive query expansion mechanism is very effective when the external collection used is much larger than the local collection and it can improve the effectiveness and robustness of query expansion.

3. Query Translation using Bilingual Dictionary

Malay-English cross-lingual IR is one of the pair of languages that are actively being researched in Malaysia [20], [21]. One of the researches is the query translation by using the bilingual dictionary where each word or phrase in source language is translated into the target language either one by one or often by several words or phrases at the same time. Translation using bilingual dictionary faces two translation problems; first on how to translate and second on how to prune alternatives. Early studies proposed using only the first translation listed in the dictionary. It is motivated by the fact that the first translation is often the most frequently used.

We carried out an experiment to test the effectiveness of the Malay-English CLIR system by using unidirectional dictionaries of Malay-English and English-Malay. A unidirectional dictionaries list the meanings of words of one language in another language. For experimentation purpose, a Malay and English document collection containing 1,446 newspaper articles is being used. From the collection, we prepared 35 Malay search queries covering a number of major events occurred in Malaysia.

First, the monolingual baseline queries were created manually using the Malay and English languages. The relevance judgments for Malay-English news collection were established manually then the term-document matrix for documents in both languages was built. Then, queries in the Malay language were automatically translated into English language and vice versa. The first translation listed in the dictionary will be selected for each translation of the term in source language. A program was built to translate the queries. The translation involves tokenization, dictionary look-up, query construction and query weighting. See Fig. 2. This experiment was evaluated using the Mean Average Precision (MAP) and Average Recall-Precision graph. The results were also being compared with the baseline for the Malay and English document collection, respectively.

4. Result and Discussion

From the experiment, it shows that the retrieval performance of the direct translation method is lower than the equivalent monolingual methods. Table 1 shows the MAP results for the baseline and cross-lingual IR using the first translation listed in the bilingual dictionary, labelled as CLIR1. English queries were translated into Malay language automatically and then being used for retrieving the Malay document collection. The MAP results for CLIR1 were lower than baseline in retrieving Malay documents. Baseline slightly outperformed the CLIR1 by 5.6%.

For retrieving the English documents, Malay queries were then translated into English language. As shown in Table 2, the MAP results for cross-lingual IR using CLIR1 also lower than the baseline result by 13.7%. The MAP result for Malay documents retrieval is higher than the English documents retrieval at the 11.2%. Table 2 and Table 3 show the average precision at 11-recall points for both the Malay and English documents retrieval.

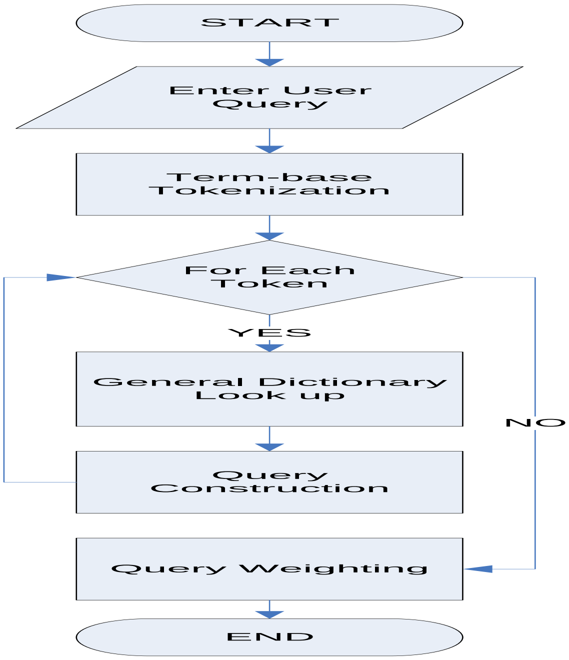


Fig. 2: The modules for user query translation

Table 1: Comparison of MAP results for baseline and CLIR1 approach

Experiment	MAP Result	
	Malay Documents	English Documents
Monolingual IR (Baseline)	0.809	0.778
CLIR1 (First Translation)	0.753	0.641

Table 2: Average precision-recall at 11-recall points for Malay monolingual IR and CLIR1

Recall	Precision	
	Monolingual IR	CLIR1
0.0	0.914	0.830
0.1	0.902	0.835
0.2	0.893	0.832
0.3	0.894	0.816
0.4	0.875	0.803
0.5	0.874	0.795
0.6	0.861	0.778
0.7	0.844	0.760
0.8	0.790	0.712
0.9	0.774	0.692
1.0	0.608	0.573
Average	**0.839**	**0.766**

Table 3: Average precision-recall at 11-recall points for English monolingual IR and CLIR1

Recall	Precision	
	Monolingual IR	CLIR1
0.0	0.866	0.759
0.1	0.846	0.742
0.2	0.844	0.720
0.3	0.836	0.707
0.4	0.826	0.696
0.5	0.802	0.663
0.6	0.794	0.647
0.7	0.766	0.636
0.8	0.725	0.588
0.9	0.703	0.572
1.0	0.573	0.450
Average	**0.780**	**0.653**

Fig. 3 and Fig. 4 show the average precision-recall in graphs form. As shown in these figures, the performance of CLIR system is lower than the monolingual IR results. The results show that if the dictionary was organized based on the most frequently used translation, retrieval system that used CLIR1 can improve the retrieval performance. One of the ways we can increase the performance is by adding all translations listed in the dictionary rather than choosing only the first translation because not all dictionaries have the same organisation. On the other hand, by including all the translations, we can obtain the query expression effect. There are two problems that have been identified in direct translation method that caused a drop in CLIR performance which are the proper names identification and translation, and compound words translation. Proper names should be identified earlier before translation event take place. Compound words handling does not applied in direct translation. As result, compound words, such as *angkasa lepas*, *kapal terbang* and *kapal selam*, were wrongly translated and it caused dropped in retrieval performance. These compound words should be translated as whole. These problems and other approaches to improve the retrieval performance will be researched in future.

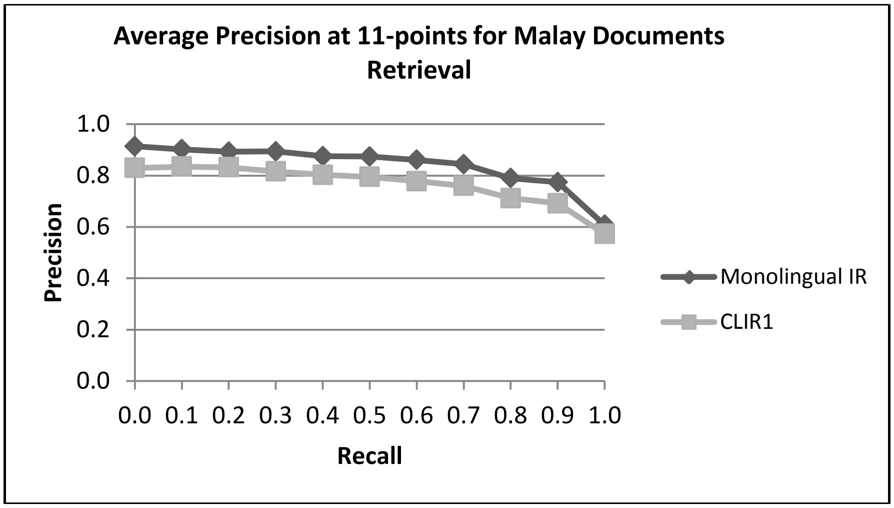


Fig. 3: The average recall-precision graph for Malay documents retrieval

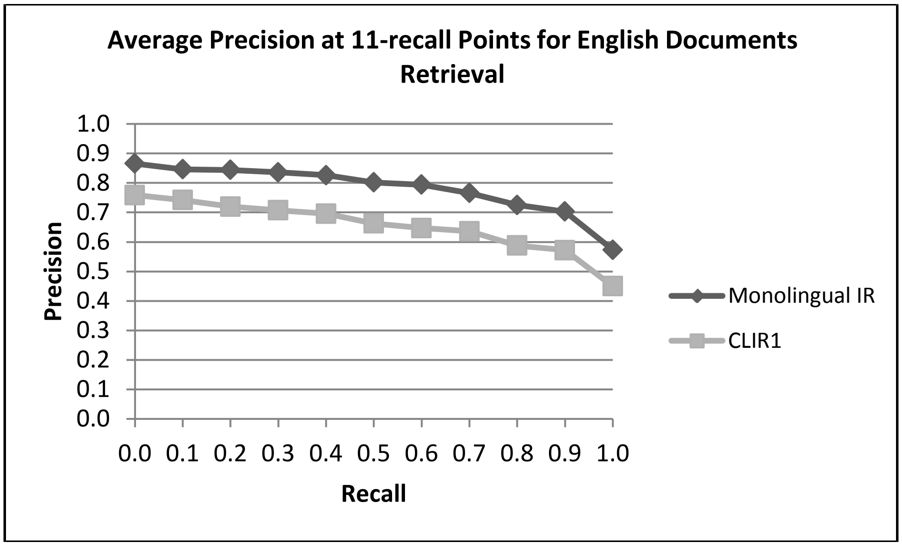


Fig. 4: The average recall-precision graph for English documents retrieval

5. Conclusion

Cross-lingual IR provides new paradigms in searching documents through myriad varieties of languages across the world and it can be the baseline for searching not only among two languages but also in multiple. Currently most of the cross-lingual research involved only top commonly used languages in the world for example English, Mandarin, Spanish and Hindi. Also, research has being done on languages that has being influenced most by the economic and commercial of a country [15]. That means in overall only a few languages have been through the development of test corpus and an extensive formal evaluation. It is unfair to only consider the cross-lingual IR research focussing only on these languages.

There are a lot of other languages that are equally important to the people which they use to search for information. There is a need for research that focuses on these languages and currently many research have been done based on the native languages of the researchers such as the Indonesian [22] and Urdu [23]. It shows that cross-lingual IR is possible to be available for every language.

When there are many languages are available to be used for searching information, there will raise the issue of multi-lingual IR. This type of IR is not restricted to only two languages but can also include as many languages as the system can. This will eventually broaden the search results that are retrieved which are relevant to the queries from the users and enhance the system's effectiveness. But on the other hand, users will have to spend time and maybe money to manually translate all the related documents to their mother-tongue language.

This report explains a description on cross-lingual IR, its issues and current methods and techniques to overcome problems for efficient and resourceful searching. This article meant for reviewing not all but some of the latest researches in the area of cross-lingual IR. As expected from the experiment done, the retrieval performance of the direct translation method is lower that the equivalent monolingual retrieval. Further investigations are needed in order to improve the effectiveness of cross-lingual retrieval focussing mainly on the Malay and English languages.

References

[1] Manning, C.D., Raghavan, P. & Schütze, H. *An Introduction to Information Retrieval*. Cambridge University Press, 2009.

[2] Abusalah, M., Tait, J. & Oakes, M. Literature review of cross language information retrieval, *World Academy of Science, Engineering and Technology*, *4,* pp. 175-177, 2005.

[3] Ren, F. & Bracewell, D.B., Advanced information retrieval, *Electronic Notes in Theoretical Computer Science, 225,* pp. 303-317, 2009.

[4] Lilleng, J. & Tomassen, S.L. Cross-lingual information retrieval by feature vectors, *Natural Language Processing and Information Systems, 4592*, pp. 229-239, 2007.

[5] Lu, C., Xu, Y. & Geva, S. Translation disambiguation in web-based translation extraction for English-Chinese CLIR, *Proceedings the 2007 ACM Symposium on Applied Computing*, pp. 819-823, 2007.

[6] Pirkola, A., Hedlund, T., Keskustalo, H. & Jarvelin, K., Dictionary-based cross-language information retrieval: Problems, methods, and research findings, *Information Retrieva*l, 4(3), pp. 209-230, 2001.

[7] Lee, C.-J., Chang, J.S. & Jang, J.-R.S., Alignment of bilingual named entities in parallel corpora using statistical models and multiple knowledge sources, *ACM Transactions on Asian Language Information Processing (TALIP), 5(2),* pp. 121-145, 2006.

[8] Chinchor, N.A., Overview of MUC-7/MET-2, *Proceedings of the 7th Message Understanding Conference (MUC-7)*, 1997.

[9] Lin, C.-C., Wang, Y.-C., Yeh, C.-H., Tsai, W.-C. & Tsai, R.T.-H. Learning weights for translation candidates in Japanese–Chinese information retrieval, *Expert Systems with Applications, 36(4)*, pp. 7695-7699, 2009.

[10] *Wikipedia*. Retrieved October 25, 2010, from: http://www.wikipedia.org.

[11] Nguyen, D., Overwijk, A., Hauff, C., Trieschnigg, D., Hiemstra, D. & de Jong, F. WikiTranslate: Query translation for cross-lingual information retrieval using only Wikipedia, *Evaluating Systems for Multilingual and Multimodal Information Access, 5706*, pp. 58-65, 2009.

[12] Gao, W., Niu, C., Nie, J-Y., Zhou, M., Hu, J., Wong, K-F., et al., Cross-lingual query suggestion using query logs of different languages, *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR'07*, pp. 463-470, 2007.

[13] Pourmahmoud, S. & Shamsfard, M. Semantic cross-lingual information retrieval, *Proceedings of the 23rd International Symposium on Computer and Information Sciences, ISCIS 2008*, 2008.

[14] Shamsfard, M., Nematzadeh, A. & Motiee, S. ORank: An ontology based system for ranking documents, *International Journal of Computer Science, 1(3)*, pp. 225-231, 2006.

[15] Gey, F.C., Kando, N. & Peters, C., Cross-language information retrieval: the way ahead, *Information Processing and Management, 41(3)*, pp. 415-431, 2005.

[16] Efthimiadis, E.N., *Query expansion*. In M.E. Williams (Ed.), Annual Review of Information Systems and Technology (ARIST), 31, pp. 121–187, 2006.

[17] White, R.W. & Marchionini, G., Examining the effectiveness of real-time query expansion, *Information Processing & Management, 43(3)*, pp. 685-704, 2007.

[18] Han, L. & Chen. G., HQE: A hybrid method for query expansion, *Expert Systems with Applications, 36(4)*, pp. 7985-7991, 2009.

[19] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. & Riedl, J., GroupLens: An open architecture for collaborative filtering of netnews, *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, pp. 175-186, 1994.

[20] Rais, N.H., Abdullah, M.T. & Kadir, R.A., Query translation architecture for Malay-English for cross-language information retrieval system, *Proceedings of the 4th International Symposium on Information Technology, ITSIM '10*, pp. 990-993, 2010.

[21] Abdullah, M.T., Monolingual and cross-language information retrieval approaches for Malay and English language document, PhD dissertation. Universiti Putra Malaysia, 2006 (Published).

[22] Adriani, M., Asian, J., Nazief, B., Tahaghoghi, S.M.M. & Williams, H.E., Stemming indonesian: A confix-stripping approach, *ACM Transactions on Asian Language Information Processing (TALIP), 6(4)*, pp. 1-33, 2007.

[23] Hussain, S., Gul, S. & Waseem, A., Developing lexicographic sorting: An example for Urdu. *ACM Transactions on Asian Language Information Processing (TALIP)*, 6(3), pp. 10, 2007.

CONSTRUCTION OF COMPUTATIONAL MALAY LEXICON USING AFFIXED WORDS ANALYSER

Harshida Hasmy[1], Zainab Abu Bakar[1], Fatimah Ahmad[2], Tengku Mohd Tengku Sembok[3]
[1] Faculty of Computer and Mathematical Sciences,
Universiti Teknologi MARA, Malaysia.
[2] Faculty of Science and Defence Technology,
Universiti Pertahanan Nasional Malaysia.
[3] Kulliyah of Information and Communication Technology,
International Islamic University Malaysia.
shidahasmy@yahoo.com

Abstract - This paper concerns an experiment on constructing computational Malay lexicon from Malay root word. A lexicon is a repository of words and that is also known as the backbone of any Natural Language Processing system that contains information about individual words or words usage. This lexicon creation will allow a morphologist, linguist or computational linguist to develop, generate, examine and manipulate a set of Malay lexicons which can improve many applications such as information retrieval systems with detailed information regarding the lexeme or word, word class, various words selection, grammatical function and many more. Usually Malay morphological processing or analysis would involve applications such as stemming, tagging or affixed analyser to determine the root words original word's structure. Using the same Malay morphological processing but with a different approach, a list of suitable Malay lexicon or words will automatically constructed from a single root word. The list of new words then will be categorized by its lexical class (nouns, verbs, adjectives, etc). This study will apply various types of affixes namely prefix, suffix, circumfix and infix on Malay root words. In this experiment, using affixed words analyser should produces list of new words constructed from one root word. As the output, from a single root word, a collection of new Malay lexicon constructed together with its classes.

1. INTRODUCTION
1.1 Overview

Malay is a language spoken by most of the people who live in Southeast Asian region such as Malaysia, Brunei, Indonesia and Singapore. The Malay language is a member of the Austronesian family of languages which is also known as "Malayo-Polynesian". Earlier when Islam arrived in South East Asia, the Arabic script (known as Jawi) was adapted to write the Malay language starting from the 14th century and replaced by the Latin alphabet (known as Rumi) in the early 20th century with the arrival of the West [15]. However, even though Malay is normally written in Rumi, Jawi is still broadly used in various fields such as legal documents, religious texts, and newspapers.

1.2 Malay Morphology

Morphological analysis is an important measure in order to obtain the Malay lexicon. Malay is an agglutinative language which is a rich morphology that performs a lot of affixation, reduplication and compounding in word formation. It creates new words by adding affixes to the root words. Malay affixation is relatively transparent and rule based A Malay word can be separated into discrete morphemes with visibly defined boundaries, including roots, prefixes, suffixes, infixes, and circumfixes. [8]. There are four basic word classes in Malay [1] which is nouns, verbs, adjectives, and function words (particles).

2. RESEARCH METHODOLOGY

The research methodology used for this study will begin with task analysis phase and ended with the result analysis.

Phase 1: Analysis of literature review

In the first phase, the analysis starts with study on the previous work done from other researchers in the same area. This literature review is focused on the discussion about Malay morphology including analyze affixation, word classes and morphographemic rules. Based on the literature [1],[2],[3] and [4] affixation includes prefixation, suffixation, circumfixaton (prefix and suffix combination) and infixation. There are a total of 21 prefixes, three suffixes, seven infixes and 27 circumfixation in Malay [1] & [4] and prefixes and suffixes are the most productive affixation process and use extensively to express grammatical relationships and to form new word. As mention in [2], some spelling variations occur when adding one native prefix or circumfix to any Malay root word. In this analysis phase, study on Malay morphographemic rules has been done to avoid any error or mistake during implementation phase where the root word will be combined with all four affixes by using the rules. Study and analysis regarding the morphological rules and word classes is crucial to increase the understanding and accuracy of the classes for each new word created and this can be obtained from [1]. In general, this phase requires some amount of time and participation from language experts is required to verify the validation of rules.

Phase 2: Implementation

In this paper, computational affixed words analyser will be used to construct a computational lexicon. The analyser is developed based on the conclusion and findings made from the literature review. In this phase, three different steps will be conducted beginning with adding any root word that can be found from dictionary [6] into the program prepared beforehand. Next, the root word will be combined with each affixation processes (prefix, suffix, circumfix and infix) by following the Malay morphographemic rules and spelling exceptions and variants to produce a new word together with its word class using the analyzer program. For this process, the standard Malay grammar rule-based approach is engaged in order to make the computer program flexible in compliant changes in the morphological rules. The analyzer must include the correct rules to avoid any errors during this process. A set of rules which defined prefixes, suffixes, circumfix (prefix-suffix pairs), and infixes are written in the following formats:

Prefix rules format: prefix+
  Example: **pe** + *latih* = pelatih,
  ter + *latih* = terlatih
Suffix rules format: +suffix
  Example: *terus* + **i** = terusi ,
  terus + **an** = terusan
Circumfix (Prefix-suffix pair) rules format: prefix+suffix
  Example: **pe**+ *lari* +**an** = pelarian,
  me+ *lari* +**kan** = melarikan
Infix rules format: +infix+
  Example: *laki* (+**el**+) = lelaki,
  gilang (+**em**+) = gemilang

Phase 3: Result analysis

This phase will evaluate the accuracy of the result by comparing each of the word against the dictionary [6] to check whether the word exists in the dictionary or not. The analyser should produce a new word together with its added affixes and word class. If the word exists in the dictionary, then it will stays in the database, otherwise more study will be perform to ensure the word can be categorized the word as a new word and or rejected word. To perform this, advice and consultation with Malay language experts will be conducted. The accepted new word then will be included into the lexicon database.

3. MALAY AFFIXATION

Affixation is one of the most common and widely used of the three morphological processes existing in Malay language. It will generate affixed words through a process which a root word may be extended by one or more affixes. In Malay, there are four types of affixes which include prefixes, suffixes, circumfixes and infixes [1]. . See Figure 1

Prefixation involve the process of adding a prefix at the left side of the root word. Common examples of prefixes are *pe-*, *pem-*, *peng-*,*me-*, *mem-*, *meng-*,*ber-*, *ter-* , *di-* and *se-* . Meanwhile, suffixation is adding a suffix to the right side of the root word and *–an*, *-kan* and *–i* are the list of suffixes involve. Circumfixation is the morphological process whereby an affix made up of two separate parts surrounds and attaches to a root at the left and right sides of the root word. A circumfix is a combination of a prefix and a suffix treated as a single morphological unit. Common circumfixes include *pe-...-an, pem-...-an, peng-...-an, per-...-an, pel-...-an,ke-...-an, me-...-kan, mem-...-kan, meng-...-kan, ber-...-kan, ber-...-an, me-...-i, men-...-i, meng-...-i,ke-...-an,di-..-i,memper-...-i,* and *diper-..-i.* Infixation is the insertion of an infix just after the first consonant of the base. There are four infixes in Malay: *-el-*, *-em-*, *er-* and *-in-*. The use of affixes will change the meaning of a word and the word class. Many initial consonants undergo mutation when affixes are added, for example from a root word *ajar* (teach). The root word *ajar* will change into a variety of new words with different meaning and word classes such as ***ajar*** = teach (verb), ***ajar***an = teachings (noun), bel***ajar*** = to learn (verb), meng***ajar*** = to teach (verb), di***ajar*** = being

taught (verb), di***ajar***kan = being taught (verb), pel***ajar*** = student (noun), peng***ajar*** = teacher (noun)
pel***ajar***an = subject (noun), peng***ajar***an = lesson, moral of story (noun) and many more.

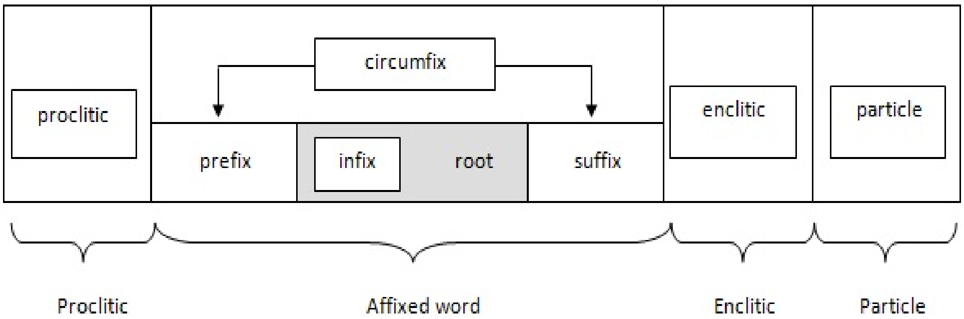


Figure 1: Affixed word with clitics (proclitic and enclitic) and particle [2]

4. WORD CLASSES

In Malay language, there are four major word classes (golongan kata) that includes Kata Nama (Noun), Kata Kerja (Verb), Kata Adjektf (Adjective) and Kata Tugas (Function Word) [1]. Each word class are divided into different sub categories, for example Kata Name (Noun) is divided into seven sub categories that is Kata Nama Am, Kata Nama Khas, Kata Ganti Nama, Kata Nama Tunggal, Kata Nama Terbitan, Kata Nama Majmuk dan Kata Nama Ganda. Same goes to Kata Kerja (Verb), where it can be separated into six sub categories which are Kata Kerja Aktif, Kata Kerja Pasif, Kata Kerja Tunggal, Kata Kerja Terbitan, Kata Kerja Majmuk, and Kata Kerja Ganda.

Even though the word can be categorized into sub categories, for this research the new constructed words will only referring into the main word classes. Later on in this research experiment, each root word that has gone through the process of adding affixes are classified into the three word class which is noun, verb and adjective according to the affixes rules involve in the process.

5. SPELLING VARIATIONS AND EXCEPTIONS

Another aspect of Malay morphology that needs to be considered is the spelling variations and exceptions. The spelling variations and exceptions only apply on some of the prefixes, circumfixes and suffixes where some of the first letter root words need to be dropped, inserted or even assimilated when combined with these affixes.

There are several spelling exceptions for prefixes prepared in [5]. Errors might occur to some words during adding affixes process if the spelling exception rules being ignored. For example, for a root word that begins with ‘***p***’ such as ***pulas*** will transform into ***memulas*** and ***pemulas*** for prefix ***‘mem’*** and ***‘pem’***, where the first letter p eliminated from the root word. Some goes to a root word ***fikir*** which will transform into ***memikir*** and ***pemikir*** after the first letter ***‘f’*** is dropped. Meanwhile, if prefix **‘me’** and **‘pe’** is added to the root word beginning with letter ***‘s’***, the letter ‘***s’*** will be dropped and replaced by ***‘ny’***. This can be applied to a root word **‘sapu’** that will transform to word ***‘menyapu’*** or ***‘penyapu’.***

Other prefixes that has the spelling exceptions are prefixes ***‘meng’*** and ***‘peng’*** which need to drop the first letter of root word that begins with letter ***‘k’***, prefixes ***‘meny’*** and ***‘peny’*** for letter ***‘s’***, and lastly ***‘men’***, ***‘pen’*** and ***‘sepen’*** for letter ***‘t’***.

There is another previous research [2] that included the morphographemic rules used to generate prefixed and circumfixed words. The rules are shown in Table 1.

TABLE 1: MORPHOGRAPHEMICS RULES FOR PREFIX AND CIRCUMFIX

Prefixes and Circumfixes	**Base**		**Morphographemics**
	Number of syllables	**Initial character**	
Group1 ber- per- ter- ber-an per-an	-	R	Deletion of initial <r>
Group 2 me- peN- peN-an	One	-	Insertion of <nge>
	More than one	a,e,i,o,u,g,h,q,x	Insertion of <ng>
		b,f,v	Insertion of <m>
		c,d,j,z	Insertion of <n>
		K	Assimilation of <k> to <ng>
		K	Insertion of <ng>
		P	Assimilation of <p> to <m>
		P	Insertion of <m>
		S	Assimilation of <s> to <ny>
		S	Insertion of <n>
		T	Assimilation of <t> to <n>
		T	Insertion of <n>

6. CONCLUSION

This paper describes the process of constructing Malay computational lexicon from Malay root word using affixed words analyser. It is hard to find similar research done on creating Malay computational lexicon from a root word. Most of the previous study or analysis would involve studies on Part Of Speech (POS), stemming, or using the affixed analyser itself to determine the root word where the words originated as well as to identify the original word’s structure. But in this study, using the affixed analyser, a list of Malay lexicon or words will automatically constructed from a single root word taken from Malay dictionary. Through this research, it produces new word that can be applied for other Malay morphological analysis. The list of lexicon stored in the database can be used for improving the process on word tagging or parsing process because all the lexicons are grouped according to their word class. This system analyser provides a starting point for future work in Malay computational lexicon and morphological analysis.

APPENDIX : AFFIXES LIST IN MALAY

AFFIXES	NOUN	VERB	ADJECTIVE
Prefix	pe-1 pe-2 peN- pem- pen- peng- penge- pel- peR- per- ke- juru-	meN- me- mem- men- meng- menge- beR- ber- teR- ter- di- mempeR- memper- dipeR- diper-	ter- se-
Suffix	-an	-kan -i	
Circumfix	pe-...-an1 pe-...-an2 peN-..-an pem-...-an pen-...-an peng-...-an penge-...-an pel-...-an peR-...-an ke-...-an	meN-...-kan me-..-kan mem-...-kan men-...-kan meng-...-kan menge-...-kan beR-..-kan ber-...-kan ber-...-an beR-..-an di-...-kan meN-..-i me-..-i mem-...-i men-...-i meng-...-i di-..-i mempeR-...-kan memper-...-kan mempeR-...-i memper-...-i ke-...-an dipeR-...-kan diper-...-kan dipeR-...-i diper-...-i	ke-...-an
Infix	-el- -er- -em-		-el- -er- -em- -in-

REFERENCES

[1] N.S Karim, F.M. Onn, and H. Musa and A. H Mahmood, (2011) “Tatabahasa Dewan Edisi Ketiga”, Dewan Bahasa dan Pustaka, Kuala Lumpur.

[2] B.R. Malancon, (2001) “Computational Analysis of Affixed Words in Malay Language”, in *8th International Symposium on Malay/Indonesia Linguistics/ SMIL8,Penang,Malaysia, 2001* pp. 1-11

[3] M.Y. Sharum, M.T. Abdullah, M.N. Sulaiman, M.A.A. Murad, Z.A.Z. Hamzah,(2010) “MALIM — A new

computational approach of Malay morphology" *in Information Technology (ITSim), 2010 International Symposium*, 2010 , vol. 2, pp. 837 - 843

[4] Y.L. Tan, (2003) "A Minimally – Supervised Malay Affix Learner", *in Proceedings of the Class of 2003 Senior Conference*; Computer Science Department, Swarthmore College, pp. 55-62,

[5] M.T. Abdullah, F. Ahmad, R. Mahmod, and T.M.T. Sembok, (2005) "A Stemming Algorithm for Malay Language", *;in Proc. CITA, 2005, pp.181-186.*

[6] Dewan Bahasa dan Pustaka, (2002) "Kamus Dewan", Edisi Ketiga, Dewan Bahasa dan Pustaka, Kuala Lumpur.

[7] G. O. Knowles, Z. M. Don,(2006) "Word Class in Malay: A Corpus Based Approach", Dewan Bahasa dan Pustaka, Kuala Lumpur.

[8] M. J. Yap, S. J. R. Liow, S. Jalil and S. S. Faizal. (2010) "The Malay Lexicon Project: A database of lexical statistics for 9,592 words", *Behavior Research Methods* , Volume 42, Number 4, pp. 992-1003.

[9] S.Sulaiman, M. Gasser , S. Kubler (2011). "Towards a Malay Derivational Lexicon: Learning Affixes Using Expectation Maximization." in *Proceedings of the* 2nd *Workshop on South and Southeast Asian Natural Language Processing, IJCNLP 2011*, pp 30-34.

[10] H. Mohamed, N. Omar, M. J. Ab. Aziz. (2011). "Statistical Malay Part-Of-Speech (POS) Tagger Using Hidden Markov Approach" *in 2011 International Conference on Semantic Technology and Information Retrieval, 2011*, pp. 231-236.

[11] Zainab Abu Bakar. (1999) ."Evaluation of Retrieval Effectiveness of Conflation Methods on Malay Documents," PhD Thesis, Universiti Kebangsaan Malaysia,1999.

[12] Z. A. Bakar and N. A. Rahman. (2003). "Evaluating The Effectiveness Of Thesaurus And Stemming Methods In Retrieving Malay Translated Al-Quran Documents", *in Proceeding of 6th International Conference On Asian Digital Libraries, 2003*, pp 653-662, Springer-verlag.

[13] F. Ahmad, A Malay Language Document Retrieval System: An Experimental Approach and Analysis, Universiti Kebangsaan Malaysia, Bangi, 1995.

[14] Musa, H., Kadir, R. A., Azman, A., & Abdullah, M. T. (2011, March). Syllabification algorithm based on syllable rules matching for Malay language. In *Proceedings of the 10th WSEAS international conference on Applied computer and applied computational science* (pp. 279-286). World Scientific and Engineering Academy and Society (WSEAS).

[15] Razak, Z., Zulkiflee, K., Salleh, R., Yaacob, M., & Tamil, E. M. (2007). A real-time line segmentation algorithm for an offline overlapped handwritten jawi character recognition chip. *Malaysian Journal of Computer Science*, *20*(2), 171.

[16] Varathan, K. D., Sembok, T. M. T., & Kadir, R. A. (2010, March). Automatic lexicon generator. In Information Retrieval & Knowledge Management,(CAMP), 2010 International Conference on (pp. 24-27). IEEE.

[17] Sembok, T. M. T., Bakar, Z. A., & Ahmad, F. (2011, July). Experiments in Malay Information Retrieval. In *Electrical Engineering and Informatics (ICEEI), 2011 International Conference on* (pp. 1-5). IEEE.

[18] Zamin, N., Oxley, A., Abu Bakar, Z., & Farhan, S. (2012). A Lazy Man's Way to Part-of-Speech Tagging. *Knowledge Management and Acquisition for Intelligent Systems*, 106-117.

[19] Tan, T. P., Xiao, X., Tang, E. K., Chng, E. S., & Li, H. (2009, August). MASS: A Malay language LVCSR corpus resource. In *Speech Database and Assessments, 2009 Oriental COCOSDA International Conference on* (pp. 25-30). IEEE.

[20] Tengku Mohd Tengku Sembok and Zainab Abu Bakar (2011). "Characteristics and Retrieval Effectiveness of n-gram String Similarity Matching on Malay Documents," in *10th WSEAS International Conference on Applied Computer and Applied Computational Science (ACACOS '11)*, Venice, 2011.

A NEW VISUAL SIGNATURE FOR CONTENT-BASED INDEXING OF LOW RESOLUTION DOCUMENTS

Danial Md Nor[1], M. Helmy Abd Wahab[2], M. Zarar M.Jenu[3] and Jean-Marc Ogier[4]

[1,2,3] *Faculty of Electrical and Electronics Engineering, Universiti Tun Hussein Onn Malaysia*
86400 Parit Raja, Batu Pahat, Johor, MALAYSIA.

[4] *L3I University of La Rochelle, 17042 La Rochelle cedix 1, FRANCE.*

danial@uthm.edu.my, helmy@uthm.edu.my, zarar@uthm.edu.my,
jean-marc.ogier@univ-lr.fr

ABSTRACT

This paper proposes a new visual signature for content –based indexing of low resolution documents. Camera Based Document Analysis and Recognition (CBDAR) has been established which deals with the textual information in scene images taken by low cost hand held devices like digital camera, cell phones, etc. A lot of applications like text translation, reading text for visually impaired and blind person, information retrieval from media document, e-learning, etc., can be built using the techniques developed in CBDAR domain. The proposed approach of extraction of textual information is composed of three steps: image segmentation, text localization and extraction, and Optical Character Recognition. First of all, for pre-processing the resolution of each image is checked for re-sampling to a common resolution format (720 X 540). Then, the final image is converted to grayscale and binarized using *Otsu* segmentation method for further processing. In addition, looking at the mean horizontal run length of both black and white pixels, the proper segmentation of foreground objects is checked. In the post-processing step, the text localizer validates the candidate text regions proposed by text detector. We have employed a connected component approach for text localization. The extracted text is then has been successfully recognized using ABBYY FineReader for OCR. Apart from OCR, we had created a novel feature vectors from textual information for Content-Based Image Retrieval (CBIR).

Keywords: Image Segmentation, Text Extraction, OCR, CBIR

Scope: Science Engineering.

1. Introduction

The presence of the text data in images and videos containing information that is useful for clearing automatically, indexing, and structuring an image. In the extraction of this information involves the detection, localization, tracking, extraction, enhancement, and recognition of text from the image provided. Nevertheless, because of diferences change the text size, style, orientation, and alignment, and low image contrast and complex background to the problem of automatic text extraction extremely challenging [1]. Although a comprehensive survey of related problems such as face detection, analysis, documents, and indexing of images & videos can be found, the problem of extracting the text information is not well explored. In this paper, we present a generic framework and methodology for automatically extract the text content of the image obtained from slide video recordings. In particular, we use the video lectures as our initial target because it can be the basis for other scenarios such as meetings and conferences. In addition to OCR, we also discuss how the unique information available in text layout. This information could be used for indexing and retrieval.

2. Related Works

Development and progress of various approaches to the extraction of text information from the image and video have been proposed for specific application, including page segmentation [2], text color extraction [3], video frame text detection [4] and content-based image or video indexing[5, 6] . Generally text-detection methods can be classified into three categories. The first one consists of connected component-based methods, which assume that the text regions have uniform colors and satisfy certain size, shape, and spatial alignment constraints. Though these methods are comparatively less sensitive to background colors, they may not differentiate the texts from the text-like backgrounds. The third one consists of the edge-based methods [7]. They found that better detection results are obtained by SVM rather than by MLP. Multi-resolution-based text detection methods are often adopted to detect texts in different scales [8]. Texts with different scales will have different features in the each sub-band. Moreover, they also made full use of the temporal redundancy to eliminate the falsely detected text regions [9]. The main contribution of this paper lies in the following three aspects: (1) it is a fast OCR method as in Fig.1, (2) it is proposed for feature extraction of textual information, and (3) it is enhancing textual information for its feature extraction.

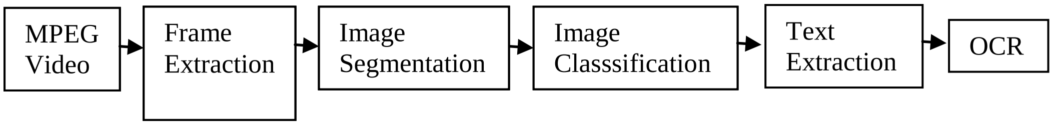


Figure 1. Flowchart of Text Extraction

The trend of attendees to take pictures of interesting slides in a presentation is becoming popular. The captured image by attendees will be more useful in academics, research and knowledge management if it can be link to the recorded presentation videos. Apart from indexing issues, the problem feature extraction of poor resolution, noisy, complex backgrounds and varying lighting conditions of the capture environment are the main issues needed to be address for better features representation.Yet another way is to provide feature vectors for low resolution image, its come from textual information.

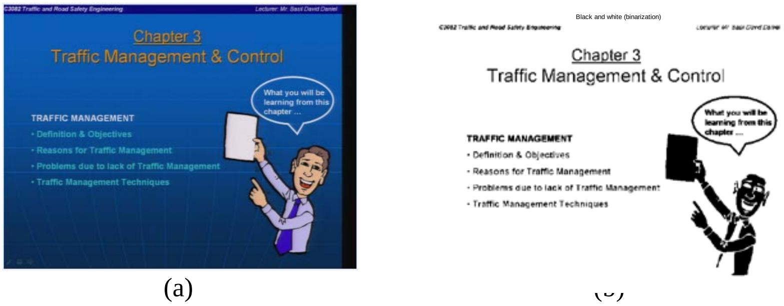
3. Text Extraction

The aim of Optical Character Recognition (OCR) is to classify optical patterns (often contained in a digital image) corresponding to alphanumeric or other characters. The process of OCR involves several steps including segmentation, feature extraction, and classification. In principle, any standard OCR software can now be used to recognize the text in the segmented frames. Due to the very low-resolution of images of those captured using handheld devices, it is hard to extract the complete layout structure (logical or physical) of the documents and even worse to apply standard OCR systems.

4. Experimental Results

4.1 Image segmentation

First of all, the resolution of each document *i.e.* both the image version of the original electronic documents and the pre-processed captured documents is checked for re-sampling to a common resolution format (720 X 540). Due to poor resolution, it is not feasible to go up to the character level as long as the adjacent characters are overlapped in the captured documents. Then, the final image is converted to grayscale and binarized using *Otsu* segmentation method for further processing [*Otsu*, 1979]. Furthermore, looking at the mean horizontal run length of both black and white pixels the proper segmentation of foreground objects is checked. For example, for the document images having dark background and light foreground, the output of the binarization is reversed *i.e.* black background (represented as 0's) and white foreground (represented as 1's) (Fig.2).



C3082 Traffic and Road Safety Engineering Lecturer: Mr. Basil David Daniel

Chapter 3

Traffic Management & Control

TRAFFIC MANAGEMENT

- Definition & Objectives
- Reasons for Traffic Management
- Problems due to lack of Traffic Management
- Traffic Management Techniques

What you will be learning from this chapter ...

(a)

Black and white (binarization)

C3082 Traffic and Road Safety Engineering Lecturer: Mr. Basil David Daniel

Chapter 3

Traffic Management & Control

TRAFFIC MANAGEMENT

- Definition & Objectives
- Reasons for Traffic Management
- Problems due to lack of Traffic Management
- Traffic Management Techniques

What you will be learning from this chapter ...

(b)

Figure 2: *Otsu* segmentation: (a) original slide document, (b) output of *Otsu* segmentation

and white pixels in the output of the *Otsu* segmentation is computed. Normally, the mean horizontal run length of the background pixels is much higher than that of foreground pixels. If the mean horizontal run length of the black pixels is comparatively higher than that of the white pixels in the binary images of the output of *Otsu* segmentation then the black and white pixels are simply swapped for the required image. Fig.2 illustrates one of such images having dark background and lighter foreground. For this particular image (Fig. 2(b)), the mean horizontal length of the black pixels is 32.3 and for white pixels it is 6.1. The image in Fig. 2(b) is corrected using this black and white run information for perfect segmentation *i.e.* black foreground and white background.

4.2 Text localization and extraction

4.2.1 Run-Length Smoothing Algorithm

The method *RLSA* is applied row-by-row and column-by-column to the above mentioned binary document images representing white pixels by 1's and black pixels by 0's. The *RLSA* transforms a binary sequence *x* into an output sequence *y* according to the rules described by *Behera* as follows [10]:

i) 1's in *x* are changed to 0's in *y* if the number of adjacent 1's is less than a pre-defined limit, *T*.

ii) 0's in *x* are unchanged in *y*.

For example, with *T* = 5 the sequence *x* is mapped into *y* , which is illustrated in Fig. 3.

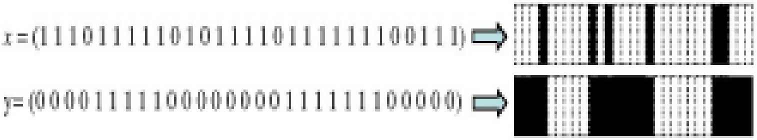
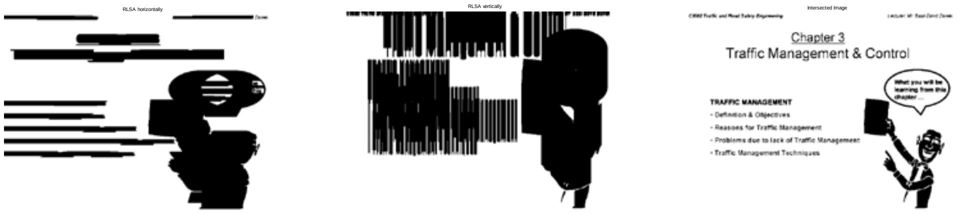

x=(11101111101011110111111100111)
y=(000011111000000000111111100000)

Figure 3: RLSA algorithm (adapted from [10])

Fig. 3 illustrates the *RLSA* algorithm in horizontal, vertical and combining output of the image in Fig. 4(c). The values of these thresholds have been evaluated and tuned using about a hundred slide images.

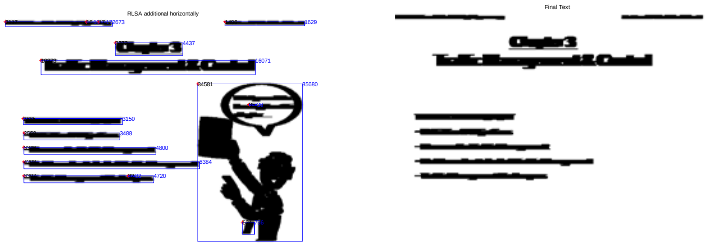

RLSA horizontally
RLSA vertically
Intersected Image
Chapter 3
Traffic Management & Control
TRAFFIC MANAGEMENT
- Definition & Objectives
- Reasons for Traffic Management
- Problems due to lack of Traffic Management
- Traffic Management Techniques
What you will be learning from this chapter ...

(a) (b) (c)

Figure 4: Output of the RLSA: (a) horizontal direction, (b) vertical direction and (c) combining both the directions, of the binary image in Fig. 4(b)

4.3 Optical Character Recognition

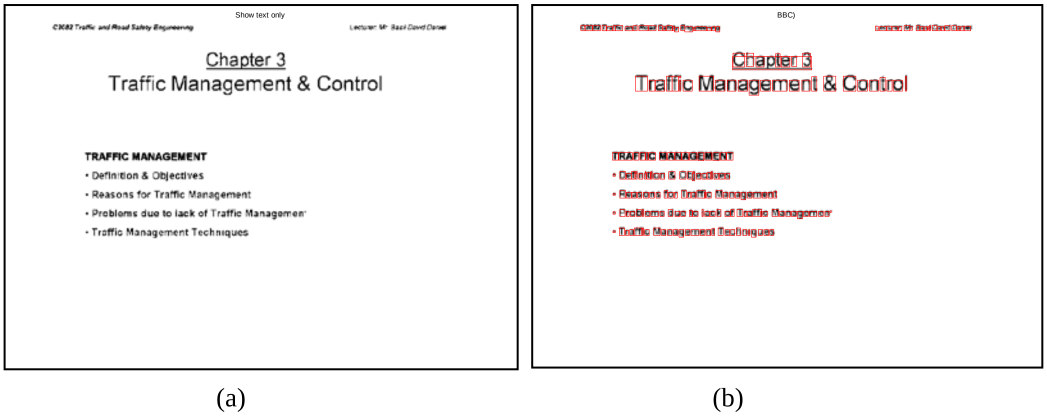
Binarization is achieved with a gray threshold value derived from Otsu's [38] method. Additional steps are done for a better binary image. Firstly, a horizontal projection profile analysis is done to detect white text on black background. If the text color is deemed to be white, the OT-region is inverted. This is necessary as most OCR software works best with black text on white background. An example of Otsu segmentation is depicted in Fig. 5.



(a) (b)

Figure 5: *Otsu* segmentation: (a) original slide document, (b) output of *Otsu* segmentation

Finally, a vertical projection profile analysis is done to discard unwanted pixel regions that are less than a calculated threshold. The final binarized image is shown in Fig. 6 (a) and its connected components is illustrated in Fig. 6(b). Connected components will be explained later. We used ABBYY FineReader 8.0 (http://finereader.abbyy.com) for OCR. We found that it was sufficient for our purpose.



Show text only

C3082 Traffic and Road Safety Engineering Lecturer: Mr Basil David Daniel

Chapter 3

Traffic Management & Control

TRAFFIC MANAGEMENT

- Definition & Objectives
- Reasons for Traffic Management
- Problems due to lack of Traffic Managemen
- Traffic Management Techniques

(a)

BBC)

C3082 Traffic and Road Safety Engineering Lecturer: Mr Basil David Daniel

Chapter 3

Traffic Management & Control

TRAFFIC MANAGEMENT

- Definition & Objectives
- Reasons for Traffic Management
- Problems due to lack of Traffic Managemen
- Traffic Management Techniques

(b)

Figure 6: The extracted text; (a) result for OCR of a binarized frame. (b) connected components.

4.3.1 Analysis of Extracted Characters

ABBYY FineReader is a good OCR software but not for low resolution document. After extracting the individual characters in a document and determining their properties, we compare them to the original set. Letters like "g", "j" & "y", which extend below the baseline are in one group, tall letters, such as "l" and "T" are in another, short characters, like "a" and "c" another, and floating characters, such as """ and "^" are in the last. Once classified, an extracted character is compared to learned characters which are in the same group. If no good match is found, the extracted character is then compared to the other original characters, regardless of group. Because we found that some characters made it past the original character recognition algorithm, we deemed it necessary to perform additional operations on poorly recognized characters. The mainly observable cause of misrecognition in our original program was linked characters as wide character. An "r" would just barely touch an "i", and the character would be recognized as an "n". To alleviate this problem, we split the character at its most narrow point. This algorithm could possibly cause problems with something like "mi"-- with a poorly scanned "m", the joined character could be broken in the middle of the "m", find an "n", and do something unpredictable with the remnants of the "m" and the "i". Another common cause of misrecognition were split characters. An "r" might be split down the middle, leaving an "l"-like figure on the left, and something incomprehensible on the right.

5. Enhancing Textual Information

In addition to OCR which was obtained from the binarized frame, we propose several other unique features that can be used to represent low resolution images. The layout shape signature can be used as a feature vector to measure the distance between two images in a query by example searching system. Here we briefly discuss the connection of two features derived from the results of the extraction of textual information Length of Sentences and the Minimum Spanning Tree of Sentences (MST).

5.1 Length of Sentences

Using text as a feature vector will result in dissimilarity between the original text and the generated ones. Hence an algorithm is proposed to classify horizontal words into three categories namely short, medium and long sentences. The sentences are determined from a number of connected components as in Fig 7. A connected component is by a set of character.

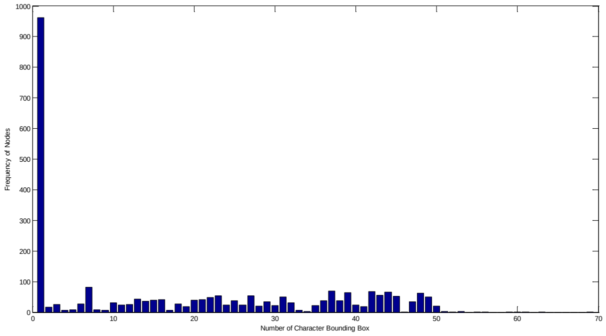
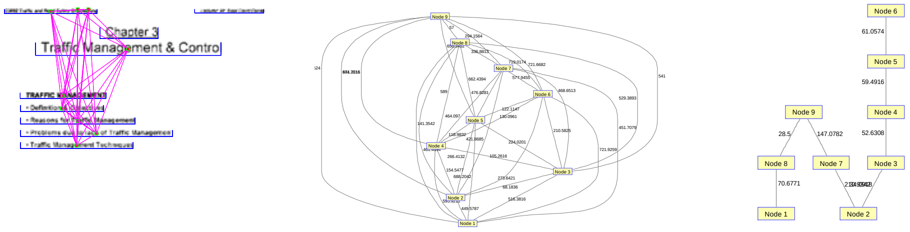


Figure 7: The histogram of classified connected components.

5.2 Minimum spanning trees

Basically a minimum spanning tree is a subset of the edges of the graph, so that there's a path form any node to any other node and that the sum of the weights of the edges is minimum. We consider each sentence as a node. Here's the minimum spanning tree of the example (Fig.8) of Prim's Algorithm by Robert C. For the example graph, here's how it would run. The resulting graph will have 8 MST as in Fig. 8(c). The details of new feature vectors proposed are; Number of Nodes / sentences, Length of MST, Average edge length (from MST), Standard Deviation of edge length (from MST), Number of small node, Number of medium node and Number of long node.



(a) (b) (c)

Figure 8: Prim's Algorithm; (a) Nodes, (b) Edges,(c) MST

6.0 CBIR Result

Figure 9 shows the good resulted obtained from a new feature vectors by using MST and Textual Profile. The unknown image from database can be recognized by measuring the Euclidean distances of two vector sets of images.

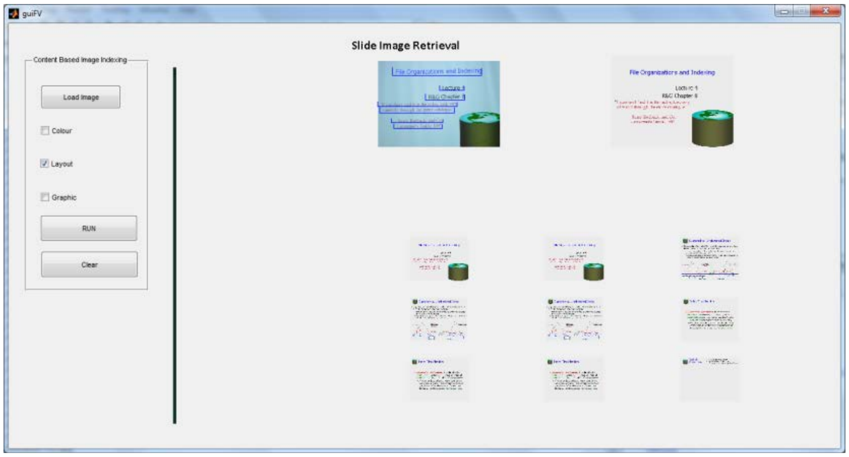


Figure 9: CBIR using Minimum Spanning Tree and Textual Profile

7. Conclusions

Image Segmentation is an important task and requires careful scrutiny. While OCR enables us to search words in the document, layout signature will be used in indexing. The good indexing scheme of an image or frame will be able to produce fast and accurate retrieval. However, the results are very dependent on the quality of OCR images or documents.

In general, the proposed algorithm can give a reliable OCR results. This allows keyword searching for the retrieval to be implemented. However, for low-resolution images, we propose a new signature taken from textual information. A new feature from textual information that we obtain as categorized nodes and the MST is unique for an image. In the future, we intend to develop Content-Based Indexing and Retrieval (CBIR) system using the obtained feature vector. We believe that through proper indexing, the proposed CBIR system is able to address the problem of low resolution document. In this paper, our focus is on scene images from the meetings and conferences. After we have an efficient CBIR, it is not difficult to make a similar system for video and image from other scenarios.

References

1. Jung, C., Q. Liu, and J. Kim, *A new approach for text segmentation using a stroke filter.* Signal Processing, 2008. **88**(7): p. 1907-1916.
2. Drivas, D. and A. Amin, *Page segmentation and classification utilizing bottom-up approach.* In *Proc. of ICDAR*, August 1995: p. 610-614.
3. Saidane, Z. and C. Garcia, *An Automatic Method for Video Character Segmentation.* 2008: p. 557–566.
4. Chen, D., J.-M. Odobez, and H. Bourlard, *Text detection and recognition in images and video frames.* Pattern Recognition, 2004. **37**(3): p. 595-608.
5. Al-Tayeche, R. and A. Khalil, *CBIR: Content Based Image Retrieval*. 2003.
6. Ameen, M.A., *Content-Based Image Search Engine*. 2004: p. 255 - 263.
7. Jung, K., K. In Kim, and A. K. Jain, *Text information extraction in images and video: a survey.* Pattern Recognition, 2004. **37**(5): p. 977-997.
8. Basu, S., et al., *Text line extraction from multi-skewed handwritten documents.* Pattern Recognition, 2007. **40**(6): p. 1825-1839.
9. Lee, S. and J. Kim, *Complementary combination of holistic and component analysis for recognition of low-resolution video character images.* Pattern Recognition Letters, 2008. **29**(4): p. 383-391.

10. Behera, A., D. Lalanne, and R. Ingold. *Enhancement of layout-based identification of low-resolution documents using geometrical color distribution*. in *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*. 2005.

TOWARDS STRUCTURE-BASED PARAPHRASE DETECTION USING DISCOURSE PARSER

Siaw Nyuk Hiong, Narayanan Kulathuramaiyer and Jane Labadin

Faculty of Computer Science and Information Technology, University Malaysia Sarawak, Malaysia
ftsm2006@yahoo.com, nara@fit.unimas.my, labadin7@gmail.com

Abstract. This study investigated the effectiveness of parsing text into syntactic structure for paraphrase detection using a discourse parser. Current state-of-the-art paraphrase detection methods have used a syntactic parser for parsing text but not a discourse parser. Furthermore, no research has been carried out to study the effect of different paraphrase mechanisms on paraphrase detection. Paraphrase instances were pre-processed using a PDTB style discourse parser and comparisons were made with results obtained from a Stanford Parser, 2-grams, 3-grams and Ferret v.3.0 (baseline) methods. Similarity scores obtained showed that text pre-processed with the discourse parser outperformed all other methods. This study initiated more future research into modeling structure-based paraphrase detection enhances by text-preprocessing using a discourse parser.

Keywords: paraphrase detection, syntactic structure, discourse parser

1 Introduction

Paraphrase detection has important application in many areas that include (a) question-answering (Ibrahim, Katz & Lin, 2003; Marsi & Krahmer, 2005), (b) text summarization (Barzilay & Lee, 2003), (c) intelligent tutoring system (McNamara et al. 2007; Rus et al., 2009) and (d) machine translation evaluation (Callison-Burch, 2008). Many researches on paraphrase detection applied a syntactic parser to pre-process text into lexico-semantic (Lintean et al. 2008; Lintean & Rus, 2010; Socher et al., 2011) or syntactic structure (Qiu et al, 2006; Wan et al. 2006; Rus et al. 2008; Mccarthy et al. 2009; Heilman & Smith, 2010; Socher et al., 2011). Lexical database like WordNet was also used to measure semantic relatedness in words or phrases (Rus et al. 2008; Das & Smith, 2009; Rus et al., 2009; Chitra & Kumar, 2010; Lintean & Rus, 2010) to improve similarity detection. Paraphrase similarity detection could be captured through various approaches as below:

(a) Machine learning approaches to model the lexical, semantic or syntactic features (Wan et al. 2006; Rus, et al., 2009; Chitra & Kumar, 2010; Socher et al., 2011).
(b) The use of threshold value as in researches by Qiu et al. (2006), Lintean et al. (2008), Lintean and Rus (2010).
(c) Graph subsumption approach by Rus et al. (2008).

Research on discourse parsers seen to be promising (Prasad et al., 2008; Elwell & Baldridge, 2008; Wellner, 2009; Pitler et al., 2009; Lin et al., 2010) but their application to paraphrase similarity detection has yet to be explored. The most current research by Socher et al. (2011) showed that syntactic tree structure was important for paraphrase detection. Since a discourse parser is able to parse a text into syntactic structure, a research could be carried out to investigate whether parsing a text with a discourse parser is equally good or better than a syntactic parser in detecting paraphrase similarity for sentences using connectives.

Even though our proposed work applied text pre-processing approach to obtain syntactic structure as in researches carried out by Qiu et al, 2006; Wan et al. (2006), Rus et al. (2008), McCarthy et al. (2009), Heilman and Smith (2010), Socher et al. (2011), it was done in a different way. First, a PDTB style discourse parser (Lin et al., 2010) was used. The parsed sentences had binary predicate-argument structure. Discourse connective is the predicate which linked two separated sentential ideas (arguments) cohesively. Second, these structural patterns captured asymmetric relationship between two separated cohesive sentential ideas (arguments) instead of only relationship between two words of a sentence as output by a syntactic parser. Stanford Parser (Klein & Manning, 2003) was used to tag all the nouns in argument 1 and 2 to capture key ideas represented. Similarity comparison could be carried out at a fine-grain level as the asymmetric relationships contain both structural and semantic information for key ideas of arguments in a sentence instead of comparing only word-to-word relationships. Third, the PDTB style discourse parser (Lin et al., 2010) produced a lower number of asymmetric relationship sets when compared to those produced by a syntactic parser. This could reduce computational cost for simple lexical matching method. Finally, the effects of different paraphrasing mechanisms in detecting paraphrase similarity were also analyzed. This kind of comparisons for detecting paraphrase similarity has not been carried out in any of the state-of-the-art research so far. Thus, our proposed work is novel in pre-processing text using a discourse parser for paraphrase similarity detection.

2 Materials and methods

2.1 Paraphrase Texts

The paraphrase texts for this case study are standard corpus for paraphrase recognition from Microsoft Research Paraphrase Corpus (Dolan et al., 2004). The selected instances covered a range of paraphrasing mechanisms that include deletion, same polarity substitution, diathesis alternation, change in the lexicalization pattern and insertion (refer Table 1). Vila, Marti and Rodriguez (2011) gave a detail description for a typology used to identify the different paraphrasing mechanisms.

Table 1. Selected instances of paraphrases for case study.

Case	Original Sentence	Case	Paraphrase	Paraphrase Mechanism
1a	The son of circus trapeze artists turned vaudevillians, O'Connor was born on Aug. 28, 1925, in Chicago and was carried onstage for applause when he was three days old.	1b	The son of circus trapeze artists turned vaudevillians, O'Connor was carried onstage for applause when he was 3 days old.	Deletion.
2a	These documents are indecipherable to me, and the fact is that this investigation has led nowhere, the lawyer said.	2b	These documents are indecipherable to me, the lawyers said, "and the fact is that this investigation has led nowhere.	Diathesis alternation.
3a	Saddam's other son, Odai, surrendered Friday, but the Americans are keeping it quiet because he's a U.S. agent.	3b	Hussein's other son, Uday, surrendered yesterday, but the Americans are keeping it quiet because he's a US agent.	Same polarity substitution.

4a	There are 625 U.N. troops in Bunia, while there are between 25,000 and 28,000 tribal fighters, from all sides, in the region.	4b	There are 625 U.N. troops in Bunia, while there are between 25,000 and 28,000 tribal fighters in the region, with thousands of them deployed in and around Bunia.	Deletion. Insertion.
5a	Don Asper called the attack "bothersome," before he and his wife contacted the firm's Web site provider to replace the vandalized page.	5b	In a telephone interview, Don Asper called the attack "bothersome," before he and his wife contacted the firm's web site provider to have the vandalised page replaced.	Insertion. Diathesis alternation.
6a	Kingston also finished with 67 after producing six birdies on the back nine.	6b	South Africa's James Kingston is also on five under after blitzing six birdies on his back nine.	Change in the lexicalization pattern.
7a	Beaumont said he doesn't expect a rapturous welcome for Jackson if he goes ahead with his visit to England.	7b	Mark Beaumont, a staff writer at music magazine NME, said he doesn't expect a rapturous welcome for Jackson if he goes ahead with his visit to England.	Deletion.

2.2 PDTB (Penn Discourse Treebank) Style Discourse Parser

PDTB (Prasad et al., 2008) annotate explicit discourse relation via lexical element known as discourse connectives. As for implicit discourse relation, it is annotated via sentence adjacency. Discourse connectives are grouped into three syntactic classes: (a) subordinating conjunctions, (b) coordinating conjunctions and (c) discourse adverbials.

A binary predicate-argument structure is used to represent PDTB discourse relation. Discourse connective is the predicate in which the two text-spans are the arguments. Argument 2 is the syntactically attached text span to the connective. Many researches using PDTB to study discourse relation and its arguments have been carried out (Wellner & Pustejovsky, 2007; Elwell & Baldridge, 2008; Pitler & Nenkova, 2009; Wellner; 2009). Lin et al. (2010) has developed a first end-to-end PDTB style discourse parser (using a parsing algorithm that connects all sub-tasks into a single pipeline) that overcome the limitations of methods developed by earlier researchers. The single pipeline of this parser is made up of (a) connective classifier, (b) argument labeler, (c) explicit classifier, (d) non-explicit classifier and (e) attribution span labeler with two sub-components: argument position classifier and argument extractor. The parser can accurately identify, label arguments and attribution spans of a discourse relation.

2.3 Word N-grams Overlap

N-grams is a set-theoretic containment based method for similarity measurement for a given n-word sequence. A document can be represented as a string of n-gram tokens before being compared (Broder, 1998). The n-gram tokens are the elements for a set which is the document. Let A and B represents two documents. Then, S(A) and S(B) will represent the sets of n-grams respectively. The five different similarity measures are as (1) to (5) below:

$$\text{simple} = |S(A) \cap S(B)| \tag{1}$$

$$\text{Dice} = \frac{2|S(A) \cap S(B)|}{|S(A)| \cup |S(B)|} \tag{2}$$

$$\text{Jaccard} = \frac{|S(A) \cap S(B)|}{|S(A) \cup S(B)|} \tag{3}$$

$$\text{overlap} = \frac{|S(A) \cap S(B)|}{\min(|S(A)|,|S(B)|)} \tag{4}$$

$$\text{Cosine} = \frac{|S(A) \cap S(B)|}{|S(A)^{\frac{1}{2}}| \times |S(B)^{\frac{1}{2}}|} \tag{5}$$

2.4 Method

Paraphrase instances were parsed using PDTB style discourse parser (Lin et al., 2010). The parsed texts consist of discourse connectives (DC) tagged arguments (ARG). In order to reduce the dimension of the DC-ARG relation, only content words (CW) consisted of noun (noun, plural noun, proper noun) in the arguments were extracted. These DC-CW-ARG were the predicate-argument tuples that represented lexical and syntactic dependencies among words of the DC-ARG for the comparison of similarities between texts. DC represented the predicate whereas CW-ARG represents the argument. Text A and Text B were thus reduced into strings of DC-CW-ARG tags to capture the underlying lexical and syntactic dependencies among words representation of the texts. Figure 1 showed the methodology used.

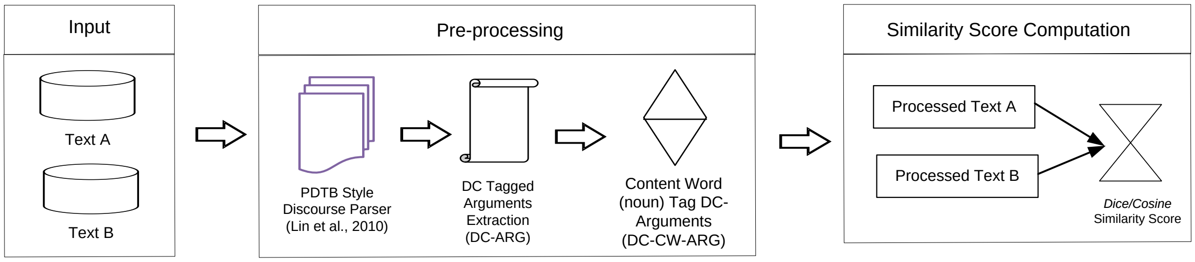

Input
Text A
Text B
Pre-processing
PDTB Style Discourse Parser (Lin et al., 2010)
DC Tagged Arguments Extraction (DC-ARG)
Content Word (noun) Tag DC-Arguments (DC-CW-ARG)
Similarity Score Computation
Processed Text A
Processed Text B
Dice/Cosine Similarity Score

Fig. 1. Case study using PDTB style discourse parser.

For the purpose of comparison with text pre-processed using a syntactic parser, Stanford Parser (http://nlp.stanford.edu:8080/parser/index.jsp) was used. Dependency relations of this parsed text consisted of words constituents with grammatical relations which become the tuples. These grammatical relations were the extracted textual relation which consisted of the triplets: name of the relation, governor and dependent. Wan et al. (2006) indicated that dependency relationships encode very similar information to bigrams. The methodology of this study was also applied to pre-processed test cases using n-gram. The only difference between the three approaches (PDTB style discourse parser, Stanford Parser, n-gram) using this methodology was in their text pre-processing. Overlap of word n-gram was used as the similarity measure for this case study. *Dice* similarity measure (symmetric score) is used since both sets representing Text A and Text B are almost of similar size. *Cosine* similarity measure (asymmetric score) is included for the purpose of comparison since it is not influence by the size of the sets. Ferret v.3.0 (2006) was used as the baseline similarity measure.

3 Results

Baseline similarity scores using Ferret v.3.0 have values ranging from the lowest score of 0.04 for test case 6 to highest score of 0.63 for test case 7. Other test cases have value around 0.4 to slightly above 0.5. Table 2 showed the scores for baseline method.

Table 2. Baseline Similarity Score.

Case	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Ferret v.3.0	0.52	0.55	0.37	0.45	0.57	0.04	0.63

Dice similarity score has the highest value of 0.90 for discourse parser in test case 5 and the lowest value of 0.07 for 3-gram in test case 6. Similar findings were also obtained for cosine similarity score. Overall, discourse parser outperformed all other methods including the baseline method for most of the test cases. Table 3 showed the similarity scores for all the test cases.

Table 3. Test cases similarity scores.

Case	**Similarity Score**	**Discourse Parser**	**Dependency Parser**	**2-gram**	**3-gram**
1	Dice	0.89	0.70	0.73	0.61
	Cosine	0.89	0.71	0.75	0.62
2	Dice	0.86	0.74	0.81	0.68
	Cosine	0.87	0.74	0.81	0.68
3	Dice	0.70	0.78	0.54	0.43
	Cosine	0.70	0.78	0.54	0.43
4	Dice	0.77	0.73	0.72	0.68
	Cosine	0.77	0.73	0.73	0.68
5	Dice	0.90	0.75	0.69	0.65
	Cosine	0.90	0.75	0.70	0.66
6	Dice	0.67	0.08	0.20	0.07
	Cosine	0.67	0.08	0.20	0.07
7	Dice	0.86	0.83	0.74	0.70
	Cosine	0.87	0.84	0.75	0.71

4 Discussion

Overall, text pre-processed with PDTB style discourse parser gave the best similarity score result with 85.71% of the test cases having the highest score. Stanford parser ranked second in giving a slightly better similarity score when compared to 2-gram method. 3-gram gave the second lowest similarity score for all the test cases. The baseline method using Ferret v.3.0 (2006) had the lowest ranking similarity score. It was not surprising since it used trigrams for comparison which was similar to 3-gram used in the case study. 3-gram actually gave the lowest similarity score ranking when compared to the other non-baseline methods. Table 4 showed the similarity score ranking.

Table 4. Similarity score ranking.

Pre-process Method	Similarity Score Ranking				
	First (%) (Highest)	Second (%)	Third (%)	Fourth (%)	Fifth(%) (Lowest)
Discourse Parser	85.71	14.28	0	0	0
Stanford Parser	14.28	42.86	42.86	0	0
2-gram	0	42.86	57.14	0	0
3-gram	0	0	0	100	0
Ferret v.3.0 (baseline)	0	0	0	0	100

The finding showed that only text pre-processed with PDTB style discourse parser was able to outperform all other methods with significant score difference for test case 6. This indicated that parsing text with this method was more efficient in detecting paraphrase mechanism that involved *a change in the lexicalization pattern*. All methods were equally efficient in pre-processed text for test case 1, 4 and 7 that used *deletion* paraphrase mechanism. Similar finding was obtained for test case 4 and 5 that used *insertion* paraphrase mechanism. In addition, test case 2 and 5 also gave similar finding for *diathesis alternation* paraphrase mechanism. On the contrary, only Stanford Parser and PDTB style discourse parser performed equally well and better than n-grams (n={2,3}) for test case 3 that used *same polarity substitution* paraphrase mechanism. Table 3 showed the similarity scores for all the test cases. Baseline similarity scores were included for comparison (Table 2).

Test case 1, 2, 4, 5 and 7 were instances that had same propositional content and similar word form for the paraphrase pairs. Even though some word/phrase was deleted (case 1, 4 and 7), inserted (case 4 and 5) or altered (case 2 and 5), formal mapping between each pair of test case was still able to give high similarity scores. This indicated that paraphrases that had same propositional content and similar word form could be efficiently detected for text pre-processed with all the four methods used in this study, with PDTB style discourse parser gave the highest similarity score overall. Paraphrase with polarity substitution (case 3) involved word substitution with synonym. This lower similar word form of the paraphrase sentence which also reduced formal mapping between the paraphrase pair. This especially affected (lower) the similarity score for text pre-processed with n-grams (n={2,3}) since n-grams could only map similar word form efficiently. On the contrary, text pre-processed with Stanford and PDTB style discourse parser was able to capture both lexical and syntactic dependencies among words of the sentence. As a result, similar word substitution did not affect the similarity scores for these methods. However, only PDTB style discourse parser could pre-process paraphrase that had same propositional content but high different word form (case 6), efficiently. Text pre-process with PDTB style discourse parser was able to capture not only lexical and syntactic dependencies among words but also separate predicate-argument relation for argument 1 and 2 of a sentence. This characteristic enabled the identification of word/phrase insertion, deletion or swapping between predicate-argument (ARG1 & ARG2) relation. Another advantage of using PDTB style discourse parser was the reduce dimension of the predicate-argument relation tuple when compare to either the Stanford

Parser or n-gram method. Table 5 showed a comparison for some of the features describe in this section.

Table 5. Text pre-processing method features comparison.

Method Features	PDTB Style Discourse Parser	Stanford Parser	2-gram	3-gram
Tuple dimension	Low	High	High	High
Similarity Score	High	Average	Average	Low
lexical and syntactic dependencies among words tuple	√	√	×	×
Word sequence tuple	×	×	√	√
Predicate(ARG1,ARG2) relation	√	×	×	×

5 Conclusion

The case studies that have been conducted in the paper showed that PDTB style discourse parser (Lin et al., 2010) outperformed all other methods including the baseline method in its effectiveness in pre-processing text for paraphrase comparison. Besides, the lower number of asymmetric relationship sets of the proposed approach comparing to either the Stanford parser or (n={2,3}-grams) method could also reduce computational cost during N-gram overlap similarity score calculation. The Predicate(ARG1,ARG2) relation output from this approach helped to enhance detection of paraphrase which has same propositional content but high different word form (case 6). The limitation of this study is it uses only 7 paraphrase instances for similarity comparisons. Since this initial study indicated positive results, future research using more instances of paraphrases could be carried out.

References

Dolan, B., Quirk, C. & Brockett, C.. (2004). Unsupervised construction of large paraphrase corpora: Exploiting massively parallel news sources. In *Proceeding Of Coling*.

Elwell, R. & Baldridge, J. (2008). Discourse connective argument identification with connective specific rankers. In *Proceedings of the IEEE International Conference on Semantic Computing*, Washington, DC, USA, 2008.

Ferret v.3.0. (2006). School of Computer Science. University of Hertfordshire. [Accessed: 13/6/2012] Available at: http://homepages.feis.herts.ac.uk/~pdgroup/

Heilman, Michael & Smith, Noah A. (2010). Tree edit models for recognizing textual entailments, paraphrases, and answers to questions. *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, pp.1011–1019.

Klein, D. & Manning, C. D. (2003). Accurate unlexicalized parsing. *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, pp. 423-430.

Lin, Z., Ng, H. T. & Kan, M. Y. (2010). A PDTB-styled end-to-end discourse parser. *Technical report TRBB/10*, School of Computing, National University of Singapore.

Mccarthy, Philip M., Guess, Rebekah H., & Mcnamara, Danielle S. (2009). The components of paraphrase evaluations. *Behavior Research Methods, 41 (3),* pp.682-690.

Pitler, E. & Nenkova, A. (2009). Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, Singapore.

Pitler, E., Louis, A. & Nenkova, A. (2009). Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Singapore. *Proceedings of SLaTE'07 Workshop*. Farmington, Pennsylvania.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A. & Webber, B. (2008). The Penn Discourse TreeBank 2.0. In *Proceedings of 6th International Conference on Language Resources and Evaluation (LREC 2008).*

Qiu, Long., Kan, Min-Yen & Chua, Tat-Seng. (2006). Paraphrase recognition via dissimilarity significance classification. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pp.18–26, Sydney, Australia, July. Association for Computational Linguistics.

Rus, Vasile., Lintean, Mihai., Graesser, Art., & Mcnamara, Danielle. (2009). Assessing student paraphrases using lexical semantics and word weighting. *Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care: From Knowledge Representation to Affective Modelling* Pages 165-172

Socher, R. Huang, E.H., Pennington, J., Ng, A.Y., & Manning, C.D. (2011). Dynamic pooling and unfolding recursive autoencoders for paraphrase detection. *Advances in Neural Information Processing Systems*, *24*.

Wan, S., Dras, M., Dale, R., & Paris, C. (2006). Using dependency-based features to take the parafarce out of paraphrase. In *Proceedings of the Australasian Language Technology Workshop*, pp. 131–138.