Automatic Thesaurus For Enhanced Chinese Text Retrieval

Schubert Foo, Siu Cheung Hui‡, Hong Koon Lim, Li Hui

School of Applied Science, Nanyang Technological University
Nanyang Avenue, Singapore 639798

Abstract

Asian languages like Japanese, Korean and in particular Chinese, are beginning to gain popularity in the information retrieval (IR) domain. The quality of IR systems has traditionally been judged by the system’s retrieval effectiveness, which in turn, is commonly measured by data recall and data precision. This paper proposes and describes a process for generating an automatic Chinese thesaurus that can be used to provide related terms to a user’s queries to enhance retrieval effectiveness. In the absence of existing automatic Chinese thesauri, techniques used in English thesaurus generation have been evaluated and adapted to generate a Chinese equivalent. The automatic thesaurus is generated by computing the co-occurrence values between domain-specific terms found in a document collection. These co-occurrence values are in turn derived from the term and document frequencies of the terms. A set of experiments was subsequently carried out on a document test set to evaluate the applicability of the thesaurus. Results obtained from these experiments confirmed that such an automatic generated thesaurus is able to improve the retrieval effectiveness of a Chinese IR system.

Keywords Chinese information retrieval, automatic thesaurus generation, co-occurrence analysis, retrieval effectiveness, data recall, data precision.

Introduction

It is a well-known fact that Chinese text is a sequence of ideological representation with no word delimiters. Thus, unlike the English and other European language, identifying words out of the Chinese text becomes a difficult task. Therefore, whenever Chinese text processing is concerned, an extra word extraction step, known as word segmentation process is required. This segmentation process will be used for all Chinese text processing related fields such as machine translation, natural language processing and information retrieval.
In general, there are two main groups of basic approaches used for word segmentation, namely, character-based and word-based approaches. In Chinese IR, character-based segmentation approaches, especially the bigram approaches (such as pure bigram and overlapping bigram) have gained widespread acceptance by many researchers (Kwok, 1997a; Kwok, 1997b; Smeaton, 1996; Tong, 1996). When employed in a Chinese IR system, the word-identification accuracy of these segmentation approaches will in turn affect the retrieval effectiveness of the system.

The retrieval effectiveness of text retrieval or any other IR systems is determined by two commonly used measures of data recall and data precision. There are a number of ways to improve the retrieval effectiveness of these systems. One of the most effective and commonly used technique is query expansion. Query expansion relies on an external aid in the form of a thesaurus to suggest terms for the user to alter his query. Thus, users are able to select related terms, broader or narrower terms to alter and refine the search.

This paper examines existing automatic thesaurus generation techniques and adapts them to the Chinese language through a proposed process. This is necessary since there are no papers published in English that details the construction of Chinese thesauri. With the constructed thesaurus, a set of experiments is performed to gauge the applicability and quality of the thesaurus. In these experiments, we are more concerned with the improvement made in the data recall measure in using the thesaurus. Additionally, improvements will also be investigated on the various bigram methods of word segmentation used for corpus indexing. These bigram variations are proposed by Li (Li, 1998) in her research on Chinese word segmentation accuracy.

**Review of Automatic Thesaurus Generation**

An IR thesaurus, unlike a language thesaurus (e.g. Roget’s Thesaurus), is very much domain specific in contrast to a general language thesaurus. Terms in an IR thesaurus are related to one another by some form of ordered hierarchical relationships (such as a broader term, narrower term, or related term). The objective of an IR thesaurus is to provide a common, precise, and controlled vocabulary that assists in the co-ordination of the indexing and retrieval processes, thus the reason of it being domain-specific. Many manually constructed thesauri exist to cover a range of vast subject areas. These thesauri are subjectively and manually-intensive constructed, and are regularly updated in view of new knowledge and advancement.

In our research, we have arbitrarily chosen an economics domain for study. Although a Chinese economics thesaurus is likely to exist, we are interested not to employ such a thesaurus but to derive a generic process for an automatic thesaurus generation that can be applied to other subject domains. Failing to locate any relevant English publications on Chinese automatic thesaurus generation, we resorted to examining the large pool of related English literature on this subject in order to study and extract relevant techniques that can be adapted for the Chinese thesaurus construction.

Frakes and Yates (Frakes, 1992) outlined three phases for the development of an automatic thesaurus:
1. Construction of vocabulary
   - Normalisation and selection of terms
   - Phrase construction depending on the co-ordination level desired

2. Similarity computation
   - Identification of statistical associations between terms

3. Organisation of vocabulary
   - Organisation of vocabulary generally into a hierarchy based on the association computed in step (2).

The following sections assess how these phases can be applied in the construction of our Chinese economics thesaurus.

Construction of vocabulary

In vocabulary construction, the relevant aspect is term selection for the thesaurus from the document collections since techniques used for the English language, namely, stemming and normalisation cannot be applied to the Chinese language. The objective of term selection is to identify the most informative terms (that can be words or phrases) for the thesaurus vocabulary from the document collection. This document collection should be sizeable and adequately represent the subject domain.

However, since all statistical methods suggested by Frakes and Yates to determine the worth of the terms require word-related attributes for computation they will not be useful if character-based Chinese text segmentation approaches are used. This is so since words segmented by these approaches can be meaningless and useless words, or ambiguous words (that can normally be combined with other segments to form more meaningful words). Such anomalies are due to the nature of the Chinese language that is still posing severe research challenges in the fields of Chinese natural language processing and machine translation. Thus, attributes computed for these words are not all useful. Therefore, these techniques of vocabulary construction are unsuitable for our purpose.

Similarity Computation

Similarity computation is concerned with deriving a relationship (i.e. the similarity measure) between pairs of terms. Two common computations are the Cosine and Dice coefficients. However, the symmetrical nature of these two measures has shown to produce undesirable side effects and poor performance (Peat, 1991). Therefore, we will adopt an asymmetric co-occurrence analysis proposed by Chen et al (Chen, 1995) instead.

When similarity measures are obtained between all term pairs, an automatic term classification process such as single-link or complete-link classification can be used to group all terms with sufficiently large similarity measures into common classes (Salton, 1978; Salton, 1989). These term-classification processes involve the use of cluster algorithms.
**Vocabulary Organisation**

Clustering algorithms are often used here to impose some structure on the vocabulary (that usually means hierarchical arrangement of term classes). A typical algorithm accepts all pair-wise similarity value corresponding to a collection of objects. It also uses values to partition the objects into clusters or classes such that objects belonging to the same class are more similar to each other than those of a different class. The use of clustering to yield such structures is optional and not a necessity since the co-occurrence analysis will provide sufficient data to allow related terms to be identified.

**Proposed Process for Automatic Thesaurus Generation**

The proposed process for automatic Chinese thesaurus generation is shown in Figure 1. This process could arguably be known as a semi-automatic process as human intervention is required at the initial stage to identify a preliminary dictionary and subsequently to refine this dictionary prior to the thesaurus construction. The process can be broken down into two sub-phases namely, the *Dictionary construction phase* and the *thesaurus generation phase*.

![Figure 1. Process for automatic Chinese thesaurus construction](image)

**Dictionary Construction Phase**

The purpose of constructing a dictionary is to aid in the extraction of domain-specific terms (i.e. economics terms) in the corpus for the generation of the automatic
thesaurus. In this way, all the term-related frequencies for the economics terms can be calculated accurately. The dictionary is constructed in the following three steps:

**Term Selection.** This is concerned with identifying a collection of terms for the preliminary dictionary. Domain-specific dictionaries are usually available for a large range of subject areas. These are special dictionaries that contain an almost exhaustive set of words associated with the subject domains. In the research, the Jing-Ji-Da-Ci-Hai, Economics Terminology Dictionary is used. This is a relatively large dictionary that spans four volumes and contains over 5000 economics terms. Although possible, it is obviously infeasible (from the point of processing and computation time) and unnecessary to include all the terms into the preliminary dictionary. This is so since many of the terms are specialised terms that are rarely found in the majority of existing economics documents. In other words, a set of commonly-used terms is likely to yield a very high percentage of terms found in commonly used economics literature.

**Term Filtering.** In term filtering, the documents in the test collection are manually parsed to identify terms that are very closely related to the economical context but are not previously found in the Economics Terminology Dictionary during the term selection step. These newly identified terms are added into the dictionary. This additional process, although optional, is helpful in the sense that all the terms accepted into the dictionary can then be considered as the domain-specific lexicons and be used to identify terms in the corpus. In the case where the document set is very large, a random sampling of documents may be selected for manual parsing and term filtering.

**Term Specification/Generalisation.** This last aspect of manual intervention is to further examine the terms that have been accepted in the previous two steps to identify additional new terms or to split existing terms. For instance, when two terms can be joined together to form a more specific and useful term, then this new composite term is added into the dictionary. For example, currency and depreciation can be combined to yield currency depreciation. In the reverse process, if a term can be split to form two or more general terms, the new smaller terms are added as well. For example, bad debts reserves can be broken into bad debts and reserves. By performing term specification and generalisation, we can have some control of the co-ordination level of our final automatic thesaurus. In our experiments, we will be aiming for a higher level of co-ordination implying that we are more concerned with the identification of more specific terms. This is so since general terms, as compared to specific terms, are relatively low in discriminative value.

It should be obvious that manual intervention can be completely eliminated by omitting the term filtering and term specification/generalisation steps. Thus, term selection can be achieved by simply selecting all the terms found in the print-based dictionary and use it as the dictionary for the latter stage of thesaurus construction. This has the distinct advantage of full automation. However, the main drawback is the likelihood of a resulting thesaurus of poorer quality and usefulness although further experiments are necessary to confirm this hypothesis.
**Thesaurus Generation Phase**

With the availability of the dictionary from the previous phase, we use the techniques proposed by Chen et al (Chen, 1995) to generate the final thesaurus.

**Compute Term Frequency and Document Frequency.** To perform the co-occurrence analysis, certain frequencies related to the terms are first calculated. The dictionary is used to identify and extract terms from the full text of each document found in the corpus. During full text parsing processing, two frequencies for every lexical term are computed. The two frequencies are term frequency, $f_{ij}$, which is defined as the number of occurrences of term $j$ in document $i$, and document frequency, $df_j$, which is defined as the number of documents in the corpus term $j$ occurs.

Each economics term found in the corpus can be considered as a descriptor to the document in which it is found but the relative importance of each descriptor in representing the content of the document varies. A weight is associated with each descriptor to denote its descriptive power and the Vector Space Model (Salton, 1989) is used to represent this term-weight association. It is well known fact that simple techniques that incorporate term frequency and inverse document frequency for weight computation have proven useful (Salton, 1989). Thus, terms that appear more times in a document (term frequency) and terms that appear in fewer documents in the whole database (more specific term, inverse document frequency) should be assign a higher weight.

**Compute Term Weights.** With the available term and document frequencies, the term weights for each term based on the product of term frequency and inverse term frequency are calculated. The weight of term $j$ in document $i$, $d_{ij}$, is computed as follows:

$$d_{ij} = tf_{ij} \times \frac{N^j}{df_j \times l_j}$$  \hspace{1cm} \text{Eqn 1}

where $tf_{ij}$ = the number of occurrences of term $j$ in document $i$; $N$ = the total number of documents in the corpus; $df_j$ = the number of documents in the corpus term $j$ occurs and $l_j$ = the number of Chinese characters in term $j$.

During the specification and generalisation process carried out in the dictionary construction previously, we observed that longer Chinese words are more specific in meaning. Thus, the length of terms is taken into weight computation and longer terms are given a higher weight.

**Perform Asymmetric Co-occurrence Analysis.** As reported by Peat and Willet (Peat, 1991), if similar terms identified by symmetric co-occurrence coefficients (such as Cosine and Dice coefficients) occurred very frequently in the corpus, it will not improve the discriminatory power of the original query. This explains the finding of Jones (Jones, 1971) that states that the best retrieval results were obtained when the less frequently occurring terms were clustered and the more frequently occurring terms were left
unclustered. Therefore, in the following asymmetric "cluster function" developed by Chen and Lynch (Chen, 1992), it will readily become apparent how these more specific terms are rewarded with a higher term weight.

\[
ClusterWeight(T_i, T_k) = \frac{\sum_{i=1}^{n} d_{jk}}{\sum_{i=1}^{n} d_{ij}} \times WeightingFactor(T_i) \quad \text{Eqn 2a}
\]

\[
ClusterWeight(T_k, T_i) = \frac{\sum_{i=1}^{n} d_{kj}}{\sum_{i=1}^{n} d_{ik}} \times WeightingFactor(T_i) \quad \text{Eqn 2b}
\]

The first equation, Eqn 2a, shows the computation of the similarity weight (or cluster weight) from term \( T_j \) to term \( T_k \) while the second equation indicates the weight computed for term \( T_k \) to term \( T_j \). \( d_{ij} \) and \( d_{ik} \) are term weights calculated using Eqn 1 in the previous step. \( d_{ijk} \) represents the combined weights of both terms \( T_j \) and \( T_k \) in document \( i \) and is defined as follows:

\[
d_{ijk} = t_{ijk} \times \log \left( \frac{N}{d_{fjk}} \times l_{j} \right) \quad \text{Eqn 3}
\]

where \( t_{ijk} = \) the number of occurrences of both term \( j \) and term \( k \) in document \( i \) (the smaller number between them is chosen) and \( d_{fjk} = \) the number of documents in which terms \( j \) and \( k \) occur together.

The \( WeightingFactor() \) in the equations Eqn 2a and 2b is there to penalise the general terms in the co-occurrence analysis. It is defined as follows:

\[
WeightingFactor(T_i) = \frac{\log N}{df_k} \quad \text{Eqn 4a}
\]

\[
WeightingFactor(T_j) = \frac{\log N}{df_j} \quad \text{Eqn 4b}
\]

Terms with a higher document frequency (i.e. more general terms) will have a smaller weight factor value, which will then produce a smaller co-occurrence probability. This causes the general terms to be pushed down the co-occurrence table in which the more relevant terms (terms with the highest cluster weights) are being listed on top. Table 1 shows a sample extract of the constructed co-occurrence table used in the experiments.
With the co-occurrence table constructed, the automatic thesaurus is now ready to be used with a Chinese IR system. In using the thesaurus, query expansion can be done simply by first identifying the keywords in the query and then querying into the co-occurrence table for the keywords’ most co-occurred terms. These retrieved terms will then be presented to the users for selection. In addition to setting a fixed number of terms for display and selection, threshold values may be set to further reduce the number of terms.

Table 1. Sample extract of a co-occurrence table used in experiments

<table>
<thead>
<tr>
<th>( T_j )</th>
<th>( T_k )</th>
<th>( \text{ClusterWeight}(T_j, T_k) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>?? :</td>
<td>?? :</td>
<td>0.284025</td>
</tr>
<tr>
<td>?? :</td>
<td>?? :</td>
<td>0.186795</td>
</tr>
<tr>
<td>?? :</td>
<td>?? :</td>
<td>0.137108</td>
</tr>
<tr>
<td>?? :</td>
<td>?? :</td>
<td>0.106780</td>
</tr>
<tr>
<td>?? :</td>
<td>?? :</td>
<td>0.106780</td>
</tr>
<tr>
<td>?? :</td>
<td>?? :</td>
<td>0.106780</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>?? :</td>
<td>?? :</td>
<td>0.175374</td>
</tr>
<tr>
<td>?? :</td>
<td>?? :</td>
<td>0.153766</td>
</tr>
<tr>
<td>?? :</td>
<td>?? :</td>
<td>0.119503</td>
</tr>
<tr>
<td>?? :</td>
<td>?? :</td>
<td>0.111520</td>
</tr>
<tr>
<td>?? :</td>
<td>?? :</td>
<td>0.091073</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
<tr>
<td>?? :</td>
<td>?? :</td>
<td>0.835986</td>
</tr>
<tr>
<td>?? :</td>
<td>?? :</td>
<td>0.769143</td>
</tr>
<tr>
<td>?? :</td>
<td>?? :</td>
<td>0.650366</td>
</tr>
<tr>
<td>?? :</td>
<td>?? :</td>
<td>0.622669</td>
</tr>
<tr>
<td>?? :</td>
<td>?? :</td>
<td>0.617073</td>
</tr>
<tr>
<td>?? :</td>
<td>?? :</td>
<td>0.595709</td>
</tr>
<tr>
<td>?? :</td>
<td>?? :</td>
<td>0.503978</td>
</tr>
<tr>
<td>:</td>
<td>:</td>
<td>:</td>
</tr>
</tbody>
</table>

**System Implementation**

As opposed to developing a Chinese IR system from scratch, an existing English-language based IR system, namely, the mg (managing gigabytes) system developed by Witten et al. (Witten, 1994) was modified to support Chinese IR and automatic thesaurus generation. The mg system is a public domain full-text retrieval system with high compression capability to enable users have instant access to an indexed collection of documents that takes up a small amount of storage space as compared to the original documents. The system supports different types of documents (textual and graphical) although only the textual component is of interest to this research. The mg system comprises two sub-systems, namely, mgbuild that is responsible for compressing and indexing of the corpus, and mgquery that is used for accessing the database built by mgbuild to process users’ queries. Both these Unix-based subsystems have been
modified to support this research. In fact, the motivation for these modifications is two-
fold. First, it supports a “plug and select” paradigm to allow different Chinese
segmentation approaches to be used for indexing and retrieval. This serves as a test-bed
for studying and evaluating different segmentation approaches and their effects on IR
performance. Second, it supports the research of automatic thesaurus generation that is
the focal point of this research. Further details of the modifications to the mg system can
be found in Lim (Lim, 1999).

Corpus Collection

The corpus used for this research comprises 3 months worth of economics news
documents that are downloaded from the People’s Daily newspaper Internet Website
(URL: http://www.peopledaily.com.cn/). The original 900 or so documents were filtered
down to 811 documents after eliminating those documents that are either too short and
containing little information worth indexing, or those charts showing the previous week’s
top earners and losers of the local stock market.

A total of 400 terms from the Jing-Ji-Da-Ci-Hai, Economics Terminology
Dictionary was identified by an employee of the mainstream Chinese newspaper agency
form the initial list of words in the dictionary. Following on from the process of term
filtering and term specification/generalisation described in the previous section, a final
list of 850 terms for the thesaurus construction was derived (Lim, 1999).

Automatic Thesaurus Construction

A total of 6 different bigram segmentation approaches has been incorporated in
the modified mg system. The interest in bigram approaches stems from the fact that an
estimation of the number of Chinese words according to the number of characters by Wu
and Tseng (Wu, 1993) revealed that there are 5% one-character words, 75% two-
character words, 14% three-character words, and 6% with four or more character words
in the Chinese language. At such, it should not be surprising to find that the majority of
existing Chinese IR systems uses some form of bigram approach for word segmentation.

The bigram segmentation approaches currently supported in the system include
the pure bigram method, overlapping bigram method, and hybrid bigram method (with
and without the use of a stoplist in all three instances), making it a total of 6 approaches.
The study of these approaches to retrieval effectiveness is another aspect of our
continuing parallel research work. Interested readers can refer to Li (Li, 1998) for more
information on this research. Although all the approaches were applied in this research,
we make comparisons among only three approaches to enable the findings to be
presented in a more concise and focussed manner. The approaches included are:

1. Pure bigram method that segments a typical sentence ABCDEFG into AB, CD, EF
and G;
2. Pure bigram method with 1-character stoplist that segments a typical sentence ABCDEFG into AB, C, DE and FG if C is a stop word, or AB, C, D, EF and G if D is a stop word;
3. Hybrid bigram method with stoplist that segments a typical sentence ABCDEFGHI, which contains CDE as an economics term found in the dictionary into AB, CDE, F, GH and I if F is a stop word, or AB, CDE, F, G and HI if G is a stopword.

The overlapping bigram approaches were dropped from comparison as their performance was almost identical to that of the pure bigram approaches.

Following the process of Figure 1, an automatic thesaurus was generated for each of the segmentation approaches. Individual databases and directories were created to store the indexed results and thesauri for the subsequent experiments.

**Retrieval Effectiveness Experiments**

A set of queries used for the experiments was derived after manually reading through the complete corpus. Likewise, the documents that are deemed relevant to the queries were also identified. Time constraints and practical considerations limited the corpus size to 3 months worth of data. Such constraints are not unusual in the context of IR research. A total of 30 queries and the relevant documents for each query are derived and identified. This formed the basis for the retrieval effectiveness experiments.

![Figure 2. Typical modified mgquery dialog](image_url)
Figure 2 shows the interface for a typical query. In the first half of the output, the retrieved mode has been set to count so that only the total number of matched documents is shown. In the second half of the output, this is subsequently set to docnums to obtain a list of document numbers, cosine measure and compressed file size. The retrieved list will enable the recall and precision percentages for this query to be computed. The remaining 29 queries are processed and the averaged percentages at the various retrieval percentiles are computed as shown in Table 2.

Table 2. Recall and precision percentages at predefined percentages of retrieval

<table>
<thead>
<tr>
<th>Retrieval Percentile (%)</th>
<th># Relevant Doc Retrieved</th>
<th># Doc Retrieved</th>
<th>Recall (%)</th>
<th>Precision (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.125</td>
<td>4</td>
<td>17</td>
<td>14.8148</td>
<td>23.5294</td>
</tr>
<tr>
<td>6.25</td>
<td>8</td>
<td>33</td>
<td>29.6296</td>
<td>24.2424</td>
</tr>
<tr>
<td>12.5</td>
<td>11</td>
<td>66</td>
<td>40.7407</td>
<td>16.6667</td>
</tr>
<tr>
<td>25</td>
<td>16</td>
<td>132</td>
<td>59.2593</td>
<td>12.1212</td>
</tr>
<tr>
<td>50</td>
<td>21</td>
<td>263</td>
<td>77.7778</td>
<td>7.98479</td>
</tr>
<tr>
<td>100</td>
<td>26</td>
<td>525</td>
<td>96.2963</td>
<td>4.95238</td>
</tr>
</tbody>
</table>

Total # Relevant Documents = 27, Total # Documents Retrieved = 525

Figure 3 shows the interface with the thesaurus capability through query expansion. Users will select additional terms as necessary and re-run the query to produce a new set of retrieved results. Obviously, tabulating the percentages for the retrieval results for queries with query expansion is not so straight forward as before since different users can (and will) select different terms from the suggested list. Therefore, we enlisted the help of 10 volunteers to provide us with 10 different samples for each query. Each of these samples will be tabulated similarly as shown in Table 2. The average values of these 10 samples are obtained, following which, the average percentages of the 30 queries are computed and percentage-curves plotted. In order to facilitate this query expansion selection exercise, we listed the suggested terms from the system for each query in a questionnaire so that the volunteers will only pick as many relevant terms (or none) from the list as necessary without having to use the mgquery system itself.
Results Evaluation

Figure 4 shows the average recall at different percentiles for all the three segmentation approaches, with and without query expansion. In all cases, it is readily apparent the use of the thesaurus has significantly improved the data recall. These improvements are more significant for both the pure bigram (19.5% on average) and hybrid bigram with stoplist (17.2% on average) methods.

The precision comparisons are not presented in the same manner as the recall percentages since this will inevitably show a drop in precision in all cases. Instead, a 3-point precision average to provide an alternative evaluation on the precision percentages is used. This 3-point precision percentage refers to the precision measured at the 25%, 50% and 75% of recall values and the precision average is simply the average of these 3 precision values. These percentages are obtained from the retrieved result of the individual query and then averaged over the 30 queries. Similarly, for the results obtained
from the query expansion samples, values for the individual queries are pre-averaged over the 10 samples.

Figure 4. Average recall percentages before and after query expansion

The choice for using this 3-point precision for evaluation instead of the precision percentages at retrieval percentile is justified. For example, in a given query, 10 relevant documents are retrieved from a set of 100 documents. Suppose the 5th and 6th relevant documents are ranked 26th and 51st in the retrieved list, then in the recall-precision table at 50% retrieved percentile, recall is 50% (5/10) and precision is 10% (5/50). However, in the three-point precision table at 50% recall (i.e. 5 relevant documents), precision is 19.23% (5/26). Thus, the precision level is actually being undermined in the first case.

Table 3. 3-point precision averages before and after query expansion

<table>
<thead>
<tr>
<th>Segmentation Approach</th>
<th>3-point Precision Average</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No thesaurus</td>
</tr>
<tr>
<td>Pure bigram</td>
<td>41.2%</td>
</tr>
<tr>
<td>Pure bigram with stoplist</td>
<td>58.5%</td>
</tr>
<tr>
<td>Hybrid bigram</td>
<td>48.7%</td>
</tr>
</tbody>
</table>
Table 3 shows the 3-point precision averages for the various segmentation approaches, with and without query expansion. Apart from the pure bigram with stoplist approach that suffered a drop when the thesaurus is used, the other two methods have shown significant improvements in the level of precision. This drop in precision for the sole case of the pure bigram with stoplist approach may be directly attributed to the result of Chinese word segmentation approach itself. Expanded query terms chosen by the users, although specific as a phrase, might be split into two or more general terms during the segmentation process. These general terms, when searched against the inverted files, will retrieve many irrelevant documents containing these general terms, thereby causing the drop in precision. One way to minimise this effect is to employ the hybrid segmentation approach that recognises these terms as distinct phrases that should not be split as seen by the improved result of this approach in Table 3.

Conclusion and Future Work

A process for generating an automatic Chinese thesaurus that can be used to provide query expansion is proposed. Although still at its infancy stage, the experiments conducted on the limited corpus have indicated that such a thesaurus enhances the level of retrieval effectiveness. With the availability of the modified mg system, future work on automatic thesaurus research can proceed in three directions. First, the corpus size can be substantially increased in the order of three-fold to five-fold to generalise these findings. At the same time, this can also include other sources of economics documents thereby leading to a derivation of a set of richer related terms. Second, it is worth using the economics dictionary in its entirety and eliminate manual intervention in constructing the dictionary as a prelude to thesaurus generation, and to make comparisons between the performance of the resulting thesaurus with the current one. Finally, the uniform process of extracting, computing frequencies and co-occurrence values does not distinguish between potential synonyms (and acronyms) so that these term and document frequencies are recorded separately. This can cause inaccuracy in the computed weights of economics terms that are related to such terms. Furthermore, as they are treated as two different terms, inter-co-occurrence weights will be computed for them. This is unnecessary because they should really be treated as a single term in the thesaurus. For example, ? ? ? ? (state-owned enterprise) which has ? ? ? ? as a synonym and ? ? as an acronym should not be treated individually but assigned a composite frequency value as a whole. In this case, all related terms to this trio can be given a more accurate weight. Furthermore, any query that includes any one of the three terms can be expanded automatically to include the remaining two terms, thereby improving the level of data recall.

References