International Journal of Computer Processing of Oriental Languages

Attributes Selection in Chinese Ontology Acquisition with FCA

GAOYING CUI,QIN LU,WENJIE LI,YIRONG CHEN

Department of Computing, The Hong Kong Polytechnic University, Hong Kong, 666666, China
csgycui, csluqin, cswjli, csyrchen@comp.polyu.edu.hk

Received (December, 9, 2007)
Revised (March 17, 2008)
Accepted (March 31, 2008)

An ontology can be seen as a hierarchical description of concepts in a specific domain. One of the key issues in ontology construction is the acquisition of ontology hierachy. Manual construction of ontology by experts is time-consuming and costly, and timely update is also difficult. In automatic or semi-automatic methods, the general procedure of ontology construction is to first obtain domain specific terms and then acquire the relations among them. Terms can be obtained through terminology extraction. Relationship among terms can then be acquired to construct an appropriate ontology hierachy. Formal Concept Analysis (FCA) is an effective tool for the acquisition and visualization of ontology structure. In using FCA to acquire ontology structure, some existing methods are inclined to use a single type of attributes such as verbs or nouns or adjectives in the context of a term. This paper investigates the use of context words including nouns, verbs, adjectives, and adverbs as well as their combination. Experiments on a general corpus and a domain specific corpus show that the combined attributes of content word types gives the best performance. Contrary to the believe that verbs serve as the best single type of attributes, nouns are better to use as they have overall better average performance over the whole spectrum of threshold values. Another experiment is conducted to examine context words which are qualified as domain specific terms from a so called Core Term List which is aquired separately. The Core Term List contains terms complying with some spedific standards as qualified domain terms. Experiments show a comparable or even better performance of using the Core Term List as context than using dynamically acquired nouns. In fact the core Term List approach gives comparable result to the combined content words. Thus, simple domain knowledge aquired in priori can serve as a good alternative attribute set especially in resource limited cases.

Keywords: Ontology, Formal Concept Analysis, Core Term List, Terminology Extraction

1. Introduction

An ontology can be seen as a hierarchical description of concepts with concepts as nodes and relations between them as links. Besides the terminology extraction as the first step, the acquisition of relationships to construct the ontology hierarchy is another important part in building an ontology. Ontology can be constructed manually or semi-automatically. Manual construction is generally by experts in the domain and it can be costly and time consuming. In the semi-automatic methods, the general procedure is to obtain domain specific terms and then acquire relations between terms with human intervention or verification in each step [1]. Formal Concept Analysis (FCA) is a formal method to model abstract objects and is shown to be a useful tool for automatic acquisition of taxonomies from texts [2]. The key in the use of FCA, however, is the selection of the appropriate attribute set. Existing methods for acquiring relationships in ontology construction are inclined to use a single type of attributes such as verbs and adjectives in the context of terms for FCA to construct a domain-specific ontology [1] [2]. For example, the rationale for using verbs is that verbs links different concepts and they can thus reflect relationships between terms. However, there is no explanations in previous works on why only a single type of words are used instead of different types of words as attributes for FCA. This paper presents an investigation on the selection of different types of attributes in automatic acquisition of ontology hierachy from Chinese corpus using FCA. Experiments are conducted on two kinds of corpora. One is a general corpus and the other is an IT domain specific corpus. The two sets of experiments are conducted using different attribute sets as candidates including part-of-speech (POS) tags of content words such as verbs, nouns, adjectives, adverbs and their combinations. The actual content words are obtained through certain statistical measures in the testing corpus. Contrary to the believe that verbs serve as the best single type of attributes, nouns are better to use as they have overall better average performance over the whole spectrum of threshold values. In addition, a separate experiment is conducted on the use of a simple static domain specific Core Term List [3] containing terms complying with some specific standards as qualified context rather than dynamically acquires attributes from context. Performance evaluation shows a comparable or even better performance than the best single attributes which are dynamically acquired and even comparable to the combined content words. Thus, simple domain knowledge aquired in priori can serve as a good alternative attribute set especially in resource limited cases.

The rest of this paper is organized as follows. Section 2 introduces the basic concepts of ontology and gives an overview of FCA. Section 3 presents related

works. Section 4 presents the experiment sets designed on the two corpora for the selection of different types of attributes along with analyses of the experimental results. Section 5 concludes the paper with discussion on future directions.

2. Basic Concepts

2.1. Definitions of Ontology

The definition of ontology originally comes from philosophy. In the 1990s, ontology began to receive wide attention from the field of computer science. An ontology is defined as a formal specification of a conceptualization in [4] for knowledge sharing. [5] gave a description to distinguish a formal ontology from an informal ontology. A formal ontology is specified by a collection of names for formal concepts and relation types organized in a ***partial ordering*** by the type-subtype relation. This study takes the formal ontology definition from [5] as given below:

Definition 1: An ***ontology***, denoted by ***O***, is defined by a quadruplet, ***O = (L, D, C, R)***, where ***L*** is a specific language, ***D*** is a specific domain, ***C*** is the set of concepts and ***R*** is the set of relations between concepts.

An ontology is normally constructed for a specific language in a given domain, thus ***L*** and ***D*** are fixed. In this work, the language is Chinese, and the domain is either the general domain or computer science. Our aim is to obtain the set of concept ***C*** and to build the set of relationships ***R*** for members of concepts in ***C***.

2.2. FCA Overview

Formal Concept Analysis (FCA) is a formal method for data analysis and knowledge representation [6]. It is a useful tool to automatically acquire taxonomies or concept hierarchies from texts.

FCA is composed of two sets of data. The first one is a so called ***object set*** and the other is an ***attribute set***. FCA can help to identify a binary relationship between the data of the two sets. The relationship is used to form a formal context according to a so-called formal concept lattice which satisfies the ***partial ordering relationship***.

The definitions of formal context and formal concept in FCA are given below according to [6].

Definition 2: A ***formal context*** is a triple (*G, M, I*) where *G* is a set of *objects*, *M* is a set of *attributes*, and *I* is the relation on $G \times M$.

Definition 3: A ***formal concept*** of the context (*G, M, I*) is a pair (*A, B*), where $A \subseteq G$, $B \subseteq M$, $A' = B$ and $B' = A$,

where $A' := \{m \in M \mid (g,m) \in I, \forall g \in A\}$ and $B' := \{g \in G \mid (g,m) \in I, \forall m \in B\}$. *A* is called the *extent* and *B* the *intent* of the formal concept.

Definition 4: A ***partial ordering relationship***, denoted by $\leq$, for two formal concepts (A_1, B_1) and (A_2, B_2) , with regard to inclusion of their extents or inverse inclusion of their intents, formalized by:

$$(A_1, B_1) \leq (A_2, B_2) \Leftrightarrow A_1 \subseteq A_2 \text{ and } B_2 \subseteq B_1$$

The whole formal concept lattice satisfies the partial ordering relations. Obviously the FCA model can be used to represent ontology where the formal concepts in FCA correspond to the concept set ***C*** for a specific language ***L*** and a specific domain ***D***. The main issues in using FCA include: (1) the selection of attribute set to describe the object set after the latter has been determined, (2) the acquisition of the mapping between ***R*** and the partial ordering relationship in FCA. In the partial ordering relationship (A_1, B_1) $\leq$ (A_2, B_2), the two objects have a so called type-subtype relation. A_1 is called the ***subclass*** of A_2 and A_2 is called the ***superclass*** of A_1. Object A_1 is subclass of A_2 if and only if B_2 is included by B_1. If A_1 is the subclass of object A_2 and A_2 is also the subclass of A_1, then the relation between A_1 and A_2 is called ***equivalence relation***.

3. Related Works

Since the time when FCA model was first brought forward, different methods have been developed for using FCA into ontology relevant researches, such as ontology design and construction, ontology update and mergence, etc.. Objects used in FCA can be in different granularity. Words, sentences or even documents can be used as objects. Attributes, on the other hands, must be contextual elements which are most descriptive to these objects, such as words and phrases in the context of the objects. Then relations between objects can then be shown through the identified relationships between their attribute sets.

The work of [2] was based on the assumption that verbs pose strong selectional restrictions on their arguments. By using a dependency parser, verb-object dependencies were extracted from the contexts with the headwords as

objects and the verbs with added postfix "able" as attributes for FCA construction. Such as the object named "*apartment*", the attribute of which can be "*rentable*" where the word "*rent*" is from the context of "*apartment*".

A method for semi-automatic ontology construction and updating with the help of FCA was proposed by [7]. The objects used in this work are nouns and only adjectives are selected as attributes. The method usd in [7] generated the initial ontology structure using several objects and their attributes, such as "*river*" with attribute "*flowing*", and "*lake*" with attribute "*stagnant*". The structure was visualized using FCA and then potential objects such as "*pond*" was added into the lattice and attributes as "*natural*", "*artifical*" were also imported to entend the ontology structure until it was completed.

The work by Li [1] was the first to apply FCA to ontology construction for Chinese. Li's work focused on how to select data sources and attribute set for FCA to construct a domain-specific ontology. She proposed a framework to facilitate the manual modification of ontology and used Information Gain (IG) to select sememes from HowNet as attributes. In the comparative experiment, a large-scale general corpus was used as the data source and only verbs were taken as the attributes used in the context of the tested object terms. This was based on the assumption that verbs pose stronger selective restrictions on their arguments or for the consideration that using verbs can avoid generating complex relations.

In the work of [8], objects in FCA are actually a domain-specific texts describing domain-specific entities, such as a whole sentence "*This apartment is in a family house, parking slot, quiet street, small bedroom, big dining room, close to public transportation. Call to set up an appt.*", in an advertisement in the real estate domain, and attributes were confined to noun phrases only, such as "*family house, parking slot*". In fact, the constructed ontology show the relationships between these descriptional sentences and phrases. There is no terminology being identified in this work.

The ontologies in [9] and [10] are constructed in the similar way as in [8]. [9] developed an ontology construction system which integrated FCA with a natural language processing (NLP) module for medical documents in which formal objects are medical documents and the compound medical phrases and concepts extracted from the NLP module serve as formal attributes.The method used in [10] described a bottom-up method to merge ontologies with the assistance of FCA. The formal objects of FCA were documents and attributes are existing concepts in ontologies obtained already. Basically, the FCA model is used in both [9] and [10] for classification of documents rather than acquisition of ontology.

4. Attribute Selection

4.1. Context Words

In this work, the interest is to take terms acquired from some terminology extraction systems, and then using context words of these terms in a corpus to identify the relationships among these terms. The objective is to build ontology for acquired domain specific terms. The method is make use of the FCA model and evaluate what are the more appropriate context words to serve as attributes for onotology construction.

It is obvious that when terms are considered as objects using the FCA model, the attributes should be acquired mainly from the context of these terms. For example, to identify the attributes to describe the term "硬盘"(hard disk), it is natural to look at the sentences where the term occurs and to look for content words in the context to serve as attributes. For example, suppose the term "硬盘"(hard disk) appears in the two example sentences

(1) "在分割硬盘时允许移动硬盘分区表." (When dividing a hard disk, the partition table of hard disk is allowed to be moved.), and
(2) "可大大提高硬盘存储照片的数量." (That can enlarge the numbers of photos which can be stored in a hard disk.)".

It is natural to take the certain context words such as "分区"(partition), "存储"(store), etc. as the attributes. It is easy to see that content words play more important semantic roles in any specific context. So in principle, nouns and verbs are the best candidates. However, adjectives and adverbs can also be considered as attributes because they directly modify nouns and verbs, respectively, and further qualify them in certain aspects. This work investigates the selection of content words including nouns, verbs, adjectives, and adverbs within context windows of object terms as attribute candidates.

Selection of attributes should take into account of different factors, such as distance from an object term, frequency of co-occurrence with the object term, etc.. Because the selection of attributes are for ontology construction, the selection criterion in this work is based on statistics similar to previous works where the content words of selected types are examined based on their co-occurrences with the domain specific terms. Only words with certain statistical significance in co-occurences are considered as descriptive attributes. A context window in the range of [-5, 5] is used for any term object to obtain a list of word bi-grams co-occurences as follows.

For each term object $\boldsymbol{t_i}$ in the object term set $\boldsymbol{T}$, a triplet $\boldsymbol{<t_i, w_j, n_{ij}>}$ is used to represented the statistics of co-occurrence between a t_i and an attaibute word $\boldsymbol{w_j}$ in the context of t_i. $\boldsymbol{n_{ij}}$ indicates their co-occurring frequency. After the collection

of statistics, all the triplets are sorted in a descending order according to n_{ij}. The system has a threshold parameter ***N***. For all $<t_i, w_j, n_{ij}>$ with $n_{ij} \geq N$, w_j is considered significant context of t_i and thus is a member of the selected attribute set.

The connection between a term t_i and an attribute word w_j is then represented by a binary membership variable pair, $\boldsymbol{\mu(t_i, w_j)}$. If there exists a triplet $<t_i, w_j, n_{ij}>$ with its w_j with $n_{ij} > N$, $\mu(t_i, w_j)$ is set to 1; 0 otherwise. Then a concept lattice can be built based on $\mu(t_i, w_j)$ for all t_i in T according to form partial ordering relations based on the FCA model first introduced in [2]. Relations between terms are represented by a $<t_i, t_j>$ tuple. Then a $<t_i, t_j>$ tuple list is generated containing all the terms in the term set T. $<t_i, t_j>$ is an ordered pair of terms according to the definition of partial ordering relationship where t_i is the subclass of t_j. If both $<t_i, t_j>$ and $<t_j, t_i>$ are in the list, it means that there are an equivalence relation between t_i and t_j.

In this work, the ConExp (Concept Explorer) system[*][11] , a Java API of open source software, is used to transform concept lattice into a visualized FCA diagram. ConExp helps to visualize the generated ontology taxonomy. If the terms in T do not appear in the corpus, they will not be shown in the FCA diagram. Those terms having no statistically significant context words, they are considered as non-attribute terms and are shown in the FCA diagram as isolated nodes only.

4.2. Core Term List

The selection of context words as attributes given in Section 4.1 is based on statistics information of the context words which requires a complexed acquisition process. However, if there are certain words that are known to be domain specific, they can be used directly rather than going through the complexed selection process. For example, in the sentence “蓝牙技术是一种无线通讯协议”(Bluetooth is a wireless communication prototol), if the words “无线”(wireless), “通讯” (communication) and “协议(protocol) are known IT terms, it is natural to use then as attributes to describe the term “蓝牙”(blutooth). Therefore, if there is already a qualified lexicon, they can be used directly as descriptive attributes for FCA construction as long as they are used the the context of the term object.

In this work, a so called Core Term List is used as a separate attribute set for ontology. The Core Term List is taken directly from a core lexicon which is obtained from a separate research work[3]. According to the definition in [3] a

[*]Copyright (c) 2000-2006, Serhiy Yevtushenko and contributors.

core lexicon in a specific domain should contain the most fundamental terms used in that domain . Each entry in a core lexicon should be frequently used in its domain. Furthermore, they must have strong descriptive power for other domain terms which means other domain terms can be built based on the core lexicon items. For example, in the IT domain, 数据(data) and 系统(system) should be core lexicon items because they can be used to form many other IT terms such as 数据库(data base), 操作系统(operating system), etc.. In this work, the Core Term List has 2,471 number of term entries taken directly from the IT domain core lexicon from [3].

The way the core lexicon items are used is very similar to that given in Section 4.1. For a given term object, if a core lexicon term appears in its context window of [-5, -5], it is qualified as an attribute for FCA construction if its frequency is higher than the current given threshold *N*.

5. Experiments and Analysis

5.1. Data Set and Experiments

Two different Chinese corpora are used in this work for attribute selection and evaluation The first corpus, referred to as the ***General Corpus,*** is a 1G corpus of People's Daily from the year 1993 to 1998 segmented and tagged by Beijing University. The second corpus, referred to tha the ***PCW Corpus***, is a 52M corpus from the Chinese magazine PC World from the year 1990 to 1996, also segmented and tagged. For comparison to previous works, the object term set *T* here is the same as that used in [1] which contains 49 IT terms as listed in Appendix A. The Core Term List contains 2,471 manually verified Chinese core terms of the IT domain.

The attribute selection algorithm given in 4.1 are applied both data sets in FCA construction.. For each corpus, there are 4 experiments to consider words for each type of POS(nouns, verbs, adjectives, and adverbs) as descriptive attributes separately. There is another experiment to investigate the combination of different POS types. Experiments for taking the Core Term List are also conducted for both corpora. For those sets of context words of terms, the ones with enough frequencies will be considered as qualified attributes.

5.2. Evaluations

As the object term set *T* is fixed, the main evaluating task is the correctness of the automatically generated partial relationships between these object terms. Before evaluation is done, a subjective answer set is prepared. Two IT domain experts

are first asked to identify the partial ordering relationships for all the terms in *T* separately. The results will then be consolidated to produce the answer set agreed by both experts after a review on results that are different. The final answer set contains 146 partial ordering term pairs from *T*. For example, for the two words "软硬件(hardware-software)" and "计算机(computer)", some people may consider that "软硬件(hardware-software)" is a subclass of "计算机(computer)". Others may think that "计算机(computer)" is a hardware, thus it is a subclass of "软硬件(hardware-software)". In this work, the word "计算机(computer)" is considered in the general sense. Thus the answer set takes the pair<"软硬件(hardware-software)" and "计算机(computer)"> as a partial ordering where "计算机(computer)" is a super class of "软硬件(hardware-software)", but not the other way around.

Then for each t_i, in *T*, the lattice will be generated based on the selected attributes w_j, where $<t_i, w_j, n_{ij}>$ is a qualified entry according to Section 4. Once the lattice is generated, it can then be applied using the FCA model to generate the partial ordering relationships. The evaluation is then conducted for each partial ordering relationship $<t_i, t_j>$ in FCA if $<t_i, t_j>$ is also in in the manually prepared answer answer set for the calculation of precision and coverage. Here the term coverage is used rather than recall because the set of terms under investigation is in the given set term set ***T*** and other terms in the corpora are not evaluated. The performance of the extracted attributes is defined by the ***f-measure*** as follows:

$$f-measure = \frac{2 * precision * \mathrm{cov}rage}{precision + \mathrm{cov}erage} \times 100\% \quad (1)$$

For the purpose of reducing the influence of corpus coverage used in the experiments, we do not calculate the relationships generated by the object terms which are not occurring in the evaluated corpus., For example, "CPU", "ASCII" and "单板机(Single Board Computer)" did not appear in the PCW Corpus, thus they are not considered in the evaluation on the PCW Corpus.

5.3. Experimental Results and Analysis

5.3.1. *Results on Different Attribute Sets*

Figure 1 shows the performance of six attribute sets on the General Corpus based on different threshold values. The six attribute sets are (1) noun only, (2) verb only, (3) adjective only, (4) adverb only, (5) content word as combination, and (6) Core Term List only.

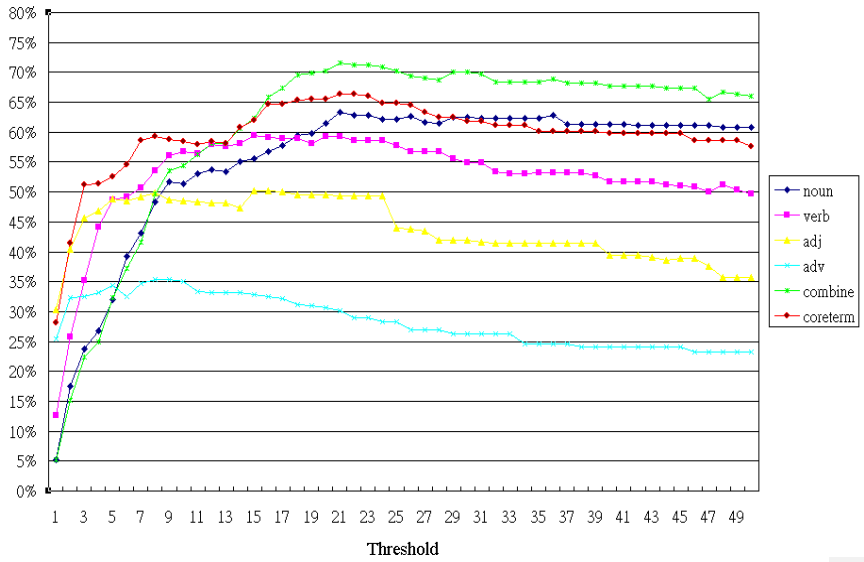

Figure 1 *F-measure* values according to threshold *N* on the General Corpus

It can be seen from Figure 1 that the highest peak point is reached by the combined content word attribute set with 71.67% in *f-measure*. The Core Term List has the second highest peak at 66.35%, yet its performance degradated to below that of nouns once the threshold reaches 30. It is interesting to note that the single noun attribute performs better than single verb attributes when *N* reaches about 18. That implies that verbs performs better than nouns with smaller threshold values. It is not difficult to understand that adjectives and adverbs are the two worst performers with f-measures of 50.18% and adverbs 35.35%, respective. From a semantic viewpoint, adjectives and adverbs are not directly associated with concepts as they are auxiliary words to nouns and verbs. So, in practice, their performances are very limited and they are not likely to be used alone.

Even though the combined content words is the best performer, its performance was the slowest to pick up. At low threshold values, it performs even worse than that of the adverbs which indicates that it requires high threshold values to pick up its discriminating power. In terms of the speed of picking up discrimination power, the Core Term list is the best performer. This is easier to explain as the Core Terms were obtained already, they do not need much of "training" to work well. Verbs also has a relatively fast pick up rate. This may

explain why many works choose verbs as attributes. But, contrary to common believe, nouns have better overall performance than verbs especially with higher threshold values.

Considering all the single attribute sets, nouns have better average performance. It indicates that nouns are more discriminative with high occurrences. It is not difficult to understand that nouns are better attributes as they are directly associated with concepts. In the General Corpus, when the threshold *N* is 21, both experiments using nouns and verbs as selected attribute set get the highest *f-measure* performance. More detailed observation shows that the words in the verb attribute set and having high co-occurrence frequency are mostly general purpose verbs, such as "用(use)", "有(have/has)", etc. The more domain specific verbs such as "识别(identify)", "查询(query)" have relatively low occurrences and thus are less likely to be used as attributes especially when threshold moves up. However, the nouns serving as attributes with high frequency are more domain domain specific such as "系统(system)", "信息(information)", "数据(data)", etc.

The experiment using the combined attribute set makes the best performances although it peaks much slower than other attribute sets. This is not surprising as this method basically takes the highest co-occurrence without considering POS tags. Experiments show that if frequency is the only consideration, in fact words of all types of content words can serve as attributes. Combination of different types of context words enhances the descriptive and discriminating power of the selected attribute set.

The combinations of adjective–noun pairs and adverb-verb are considered and evaluated. But the performances of these two combinations are no better than using nouns and verbs as single attributes. Thus we did not show their results here and they are not included in subsequent analysis. The poorer performance is most likely due to data sparsity problem.

The IT domain specific Core Term list performs well in the General Corpus and is better than the noun attributes. This is because the Core Term list contains the most fundamental and most frequently used terms used in the IT domain. Thus it has no conflict with the data in the IT domain. In fact, many core terms are frequently used in the General Corpus such as the domain specific term "图像处理(image processing)" which is concatenated by two core terms "图像(image)" and "处理(processing)".

Figure 2 shows the performance results on the PCW Corpus using the same 6 attribute set. In fact, the performance trend for all the six attribute sets are similar compared to that for the General Corpus except two obvious differences. First, the pick up trend is much less stable than that shown in Figure 1. Second, the

pick up speed is much slower. That is why this experiment has a longer range of *N* of 120 as compared to 50 in Figure 1.

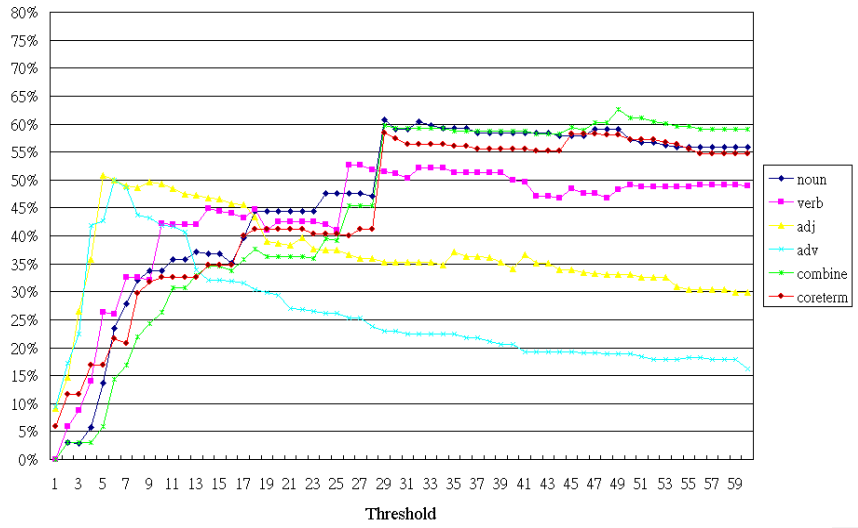

Figure 2 *F-measure* values according to threshold *N* on the PCW Corpus

As shown in Figure 2, the highest peak point is also reached by the combined attribute set with *f-measure* value of 62.61%. Then is the noun attribute set with 60.71%. The experiment considering only verbs only reaches 52.63% as its peak point followed by adjectives of 50.75% and adverbs of 50%. The orders of reaching peak points are almost the same for both corpora. The curve of the Core Term list only trails the curve of the noun attribute slightly because the Core Term list is more relevant in the IT domain corpus.

The reasons for a slower curves pick up and less stable increase can be explained by more detailed observations. The number of words that can be used as attributes in the PCW Corpus is much larger than that of the General Corpus. The average numbers of attributes one object can have in the General Corpus and the PCW Corpus are around 63 and 319, respectively. Which means that there are more relevant context words under the same threshold values. On the other hand, this also means that there are more potential noises as far as the FCA model is concerned making the system less stable. For example, when threshold is set as 1, the numbers of average combined attributes from the General Corpus and PCW corpus for one object are 500 and 2,000 respectively. In fact, this also

explains why it takes higher threshold values to weed out the attributes that are making more noise than contributing to performance.

For more detailed analyses, Figure 3 and Figure 4 show the distribution of the different types of words under different threshold values on the General Corpus and the PCW Corpus, respectively. The distribution in Figure 3 indicates that the percentage of different word types are quite stable with about 52% as verbs, 42% as nouns, and 6% as adjectives. This in a way justdifies the intuition that verbs are commonly selected as attributes. However, 52% is only slightly over the 50% mark which clearly indicates that using verbs alone is not enough. By looking at the memberships of actual words used as attributes in different threshold values, different words types are quite different. The number of noun attributes is reduced from 4,955 to 82 and when *N* is changed from 1 to 50 which is similar to that of verbs from 3,683 to 87. However, most of the domain specific nouns such as "芯片(chip)", "软件(software)" are still in the attribute set. Yet, most of the more domain relevant verbs words such as "读取(read)", "存储(store)", "开发(develop)" have disappeared. Adjectives are in the same state as verbs. This explains why the *f-measure* value of nouns remains steady when *N* increases, but the *f-measure* value of both the verbs and adjectives have a downward trend with an increased *N* value.

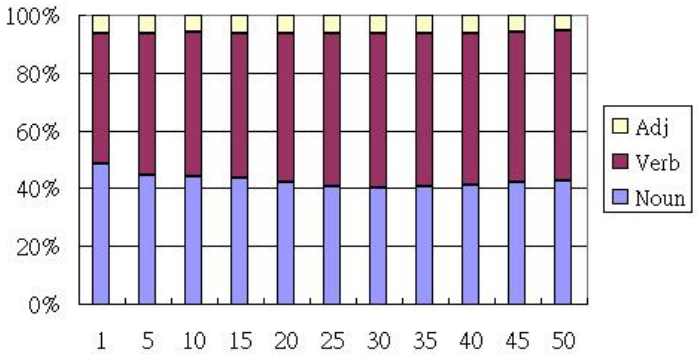

Figure 3 Distribution of different word types on the General Corpus

Figure 4 for domain specific data shows a different trend. Not only the percentage of nouns are larger, it shows a increasing trend from the beginning of 45% to more than 51%. The percentages of verbs and adjectives drops from 43% to 39%, and 10% to 8%, respectively. This is a good indication that in a domain specific corpus, nouns play a more important role and their importance cannot be overlooked.

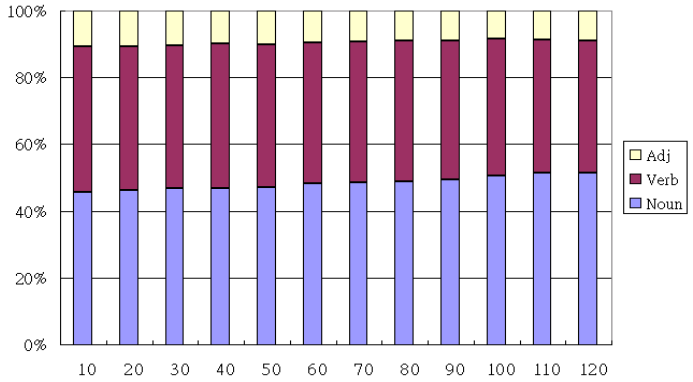

Figure 4 Distribution of different word types on the PCW Corpus

5.3.2. *Average Performances on Different Corpora*

Comparing the *f-measure* values in Figure 1 and Figure 2, it can be seen that that the performances on the IT corpus is poor than the results on the General Corpus. This is contratory to common believe that words in a domain specific corpus should be of better quality and more discriminative than words in a general corpus to describe domain specific terms. To explain this issue, more detailed analysis on precison and coverage is needed. Table 1 and Table 2 shows the average precision and coverage of the two corpura, respectively.

Corpus	Average Precision				
	Noun	Verb	Adj	Combined	Core Term
General Corpus	55.47%	47.79%	33.25%	77.11%	56.76%
PCW Corpus	63.41%	51.99%	27.46%	75.01%	62.88%

Table 1 Average precisions of both corpora

Corpus	Average Coverage				
	Noun	Verb	Adj	Combined	Core Term
General Corpus	60.50%	60.25%	67.01%	53.77%	65.54%
PCW Corpus	40.15%	45.79%	68.54%	36.23%	40.33%

Table 2 Average coverages of both corpora

Table 1 shows that the precisions of applying candidate attribute sets on PCW corpus are higher than or comparable to that of the General Corpus. But when considering the coverage to the standard answer, the General Corpus is better as shown in Table 2. That means the performance degradation of the PCW is caused by a lower coverage.

To further examine the attributes used for term identification and relationship identification, more detailed statistical analysis has been done on the two corpora using an attributes to object ratio $\boldsymbol{R_{AtoO}}$, defined by the total number of attributes divided by the total number of term objects. R_{AtoO} shows on average the number of attributes used to describe each term object. The statistical data on R_{AtoO} are shown in Figure 5 and Figure 6 for the five different attribute sets for the General Corpus and the PCW Corpus, respectively.

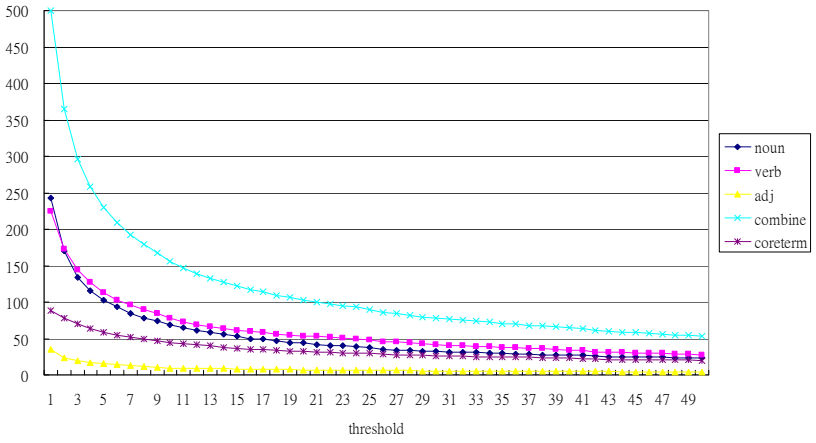

Figure 5 Attributes to object ratios on the General corpus

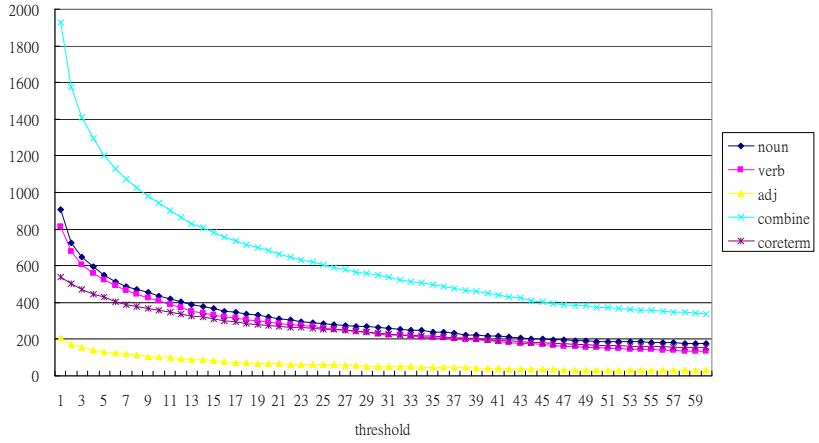

Figure 6 Attributes to object ratios on the PCW Corpus

Even though Figure 5 and Figure 6 exhibit similar trends for the 5 attributes sets examined, the scale of the figure for the PCW Corpus is much bigger. Taking the curves of the combined attribute in Figure 5 and Figure 6 as an example, the ratios are descending on both corpora. However, the maximal ratio on the General Corpus is up to 499 when the threshold value is 1. While on the PCW Corpus, the maximal ratio is up to 1930, which is about 4 times of that in the General Corpus. The best performances on both corpora are achieved when the ratio becomes less than 100. This means that when there are too many attributes, the performance is lower This is consistent for all the other attribute sets too. The below Table 3 gives a summary of $\boldsymbol{R}_{AtoO}$ for both copora.

Corpus	Average Attributes/object Ratio($\boldsymbol{R}_{AtoO}$)				
	Noun	Verb	Adj	Combined	Core Term
General Corpus	52	60	8	119	34
PCW Corpus	308	275	65	648	257
$\boldsymbol{R}_{AtoO}$-*PCW*/$\boldsymbol{R}_{AtoO}$-*General*	5.9	4.6	8.1	5.4	7.6

Table 3 Average attributes to object ratios of both corpora

It can be seen from Table 3 that on average, every term object has more attributes to describe it in the PCW Corpus. This is in a way consistent with the

common understanding that there are more context information in a domain specific corpus. However, the more attributes each term object has, the more likely that the attributes of different term objects would intersected rather than one being a subset of another making it more difficult to satisfy the required partial ordering relationship. In other words, less partial ordering relationships can be identified using the FCA model in a domain specific corpus. This explains the relatively low coverage of the PCW Corpus compared to the General Corpus. For example, in the General corpus, there is a partial ordering relationship from the term "硬盘(hard disk)" to "软硬件(hardware-software)" when using the combined attributes. But such a relationship cannot be identified in PCW corpus using the FCA analysis. This is because all the combined attributes of "软硬件(hardware-software)" from the General Corpus are included in the attribute set of "硬盘(hard disk)". But when using the PCW Corpus, there are some attributes of "软硬件(hardware-software)" not occurring in the attribute set of "硬盘(hard disk)", such as "平台(platform)" and "大型机(mainframe computer)".

This suggest that more attributes sometimes can make the construction of ontology more difficult as too much details can make the construction of a simple hierarchical structure difficult. Future work can be done to further prune out smaller details if FCA is used. Other methods such as further restrictions using syntactic chunks can also be explored.

6. Conclusion and Future Work

This paper investigates the use of context words including nouns, verbs, adjectives, and adverbs as well as their combination. Experiments on a general corpus and a domain specific corpus show that the combined attributes of content word types gives the best performance. Contrary to the believe that verbs serve as the best single type of attributes, nouns are better to use as they have overall better average performance over the whole spectrum of threshold values. Another experiment is conducted to examine context words which are qualified as domain specific terms from a so called Core Term List which is aquired separately. The Core Term List contains terms complying with specific standards as qualified domain terms. Experiments show a relative better performance of using the Core Term List as context than using dynamically acquired single attributes. In fact the Core Term List approach gives comparable result to the combined content words. Thus, simple domain knowledge aquired in priori can serve as a good alternative attribute set especially in resource limited cases.

One way to enhance the performance of the attribute set is to add position information for each selected context word. With position information, verbs as

attributes can help to discriminate a term object as an agent or theme of another one. Other attribute words can also help to determine the precedence of the objects in a relation. Another possible direction could be to loose the total inclusion rule for the partial ordering relation in the FCA construction. More comprehensive algorithms rather than simple co-occurrence frequency can also be investigated. Syntactic and semantic cues can also be used to improve the quality of ontology construction.

Acknowledgments

This work is partially supported by CERG grants (PolyU 5190/04E and PolyU 5225/05E) and a Hong Kong Polytechnic University funded project 4-Z08K..

References

[1] Sujian Li, Qin Lu, Wenjie Li, 2005, Experiments of Ontology Construction with Formal Concept Analysis. In proceedings of OntoLex2005, the 4th workshop of Ontologies and Lexical Resources, Jeju Island, South Korea, October 15, 2005, pp67-75.

[2] P Cimiano, S Staab, & J Tane, 2003, Automatic Acquisition of Taxonomies from Text: FCA Meets NLP. In Proceedings of the ATEM-2003, the PKDD/ECML'03 International Workshop on Adaptive Text Extraction and Mining. Cavtat-Dubrovnik (Croatia), September 22, 2003.

[3] Luning Ji, Qin Lu, Wenjie Li, YiRong Chen, 2007, Automatic Construction of a Core Lexicon for Specific Domain. In proceedings of ALPIT2007, the 6th International Conference on Advanced Language Processing and Web Information Technology, Luo Yang, China, August 22-24, 2007.

[4] TR Gruber, 1995, Toward Principles for the Design of Ontologies Used for Knowledge Sharing. In proceedings of International Journal of Human and Computer Studies, 43(5/6):907-928.

[5] JF Sowa, 2000, Knowledge Representation, Logical, Philosophical, and Computational Foundations. Brooks Cole Publishing Co., Pacific Grove, CA, 2000.

[6] B Ganter, & R Wille, 1999, Formal Concept Analysis. Mathematical Foundations. Berlin-Heidelberg-New York: Springer, Berlin-Heidelberg.

[7] M Obitko, V Snasel, J Smid, 2004, Ontology Design with Formal Concept Analysis. In proceedings of the CLA 2004 Workshop: Concept Lattices and Their Applications Workshop, Czech Republic, Sep 2004, pp 111-119.

Formatted: English (U.K.)

[8] HM Haav, 2003, An Application of Inductive Concept Analysis to Construction of Domain-specific Ontologies. In proceedings of the Pre-conferenceVLDB2003, the 29th International Conference on Very Large Data Bases, Los Altos, September 9-12, 2003, pp 63-67.

[9] Guoqian Jiang, K Ogasawarab, A Endoha, & T Sakuraia, 2003, Context-based ontology building support for clinical domains using formal concept analysis. International Journal of Medical Informatics (2003) 71, pp71-81.

[10] G Stumme, A Maedche, 2001, FCA-Merge: bottom-up merging of ontologies.In proceedings of 7th International Conference on Artificial Intelligence (IJCAI'01), Seattle, WA, USA, 2001, pp 225-230.

[11] SA Yevtushenko, 2000, System of data analysis "Concept Explorer" (in Russian.). In 7th National Conference on Artificial Intelligence KII-2000, (pp. 127--134). (Available at http://sourceforge.net/projects/conexp)

Appendix A

Terms		
ASCII	CPU	编程(programming)
操作系统(operating system)	存储器(storage)	存取(storing)
大型机(mainframe computer)	单板机(Single Board Computer)	电脑(computer)
电子商务(e-business)	电子邮件(email)	调制解调器(modem)
读取(reading)	服务器(server)	工作站(workstation)
光标(cursor)	硅谷(Silicon Valley)	缓存(cache)
回车(return)	寄存器(register)	计算机(computer)
计算机辅助(computer aided)	计算机化(computerization)	计算中心(computing centre)
监视器(monitor)	兼容机(compatible machine)	键盘(keyboard)
解码(decoding)	空格(space)	联机(online)
模式识别(pattern recognition)	内存(memory)	屏幕(screen)
软硬件(hardware and software)	数字计算机(digital computer)	图标(icon)
微处理机(microcomputer)	微处理器(microprocessor)	微电脑(microcomputer)
微机(microcomputer)	微型机(microcomputer)	硬磁盘(hard disk)
硬盘(hard disk)	中央处理器(CPU)	终端(terminal)
终端机(terminal)	主板(mainboard)	字节(byte)
总线(bus)		