Experiments of Ontology Construction with Formal Concept Analysis

Sujian Li
Institute of Computational Linguistics
Peking University
Beijing, China, 100080
lisujian@pku.edu.cn

Qin Lu
Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
csluqin@comp.polyu.edu.hk

Wenjie Li
Department of Computing
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
cswjli@comp.polyu.edu.hk

Abstract

An ontology can be seen as a system of concepts. Formal Concept Analysis (FCA) is a formal method to model abstract objects. In this paper we address the issue of how to select data sources and attribute set so that FCA can be used in constructing a domain-specific ontology. Two experiments are designed using two different kinds of data sources. One uses the HowNet lexicon and the other uses a large-scale corpus. The corresponding attributes including a set of sememes and a set of context verbs are used as attributes, respectively. The experiments are made to gain insight as to what are the important issues in ontology construction and trade-off in efficiency. In can be seen that in manual ontology construction, both the choice of sememes and their granularity are important as well as the correct mapping of the terms to the set of attributes. On the other hand, the methods to select the types and related context words in the corpus are important. Pure statistical method without regards to syntax and semantics would result in an ontology which is difficult to interpret. It is also shown that the use of visual tool such as FCA model is very helpful in building a good ontology.

1 Introduction

An ontology can be seen as a system of concepts in a specific domain. How to construct an ontology is a non-trivial task. The ontology are normally constructed either manually or semi-automatically. In the manual methods, the concept architecture is manually constructed by experts who had to look through all kinds of dictionaries to obtain the indivisible concept atoms. For example, HowNet is constructed by experts with a long and tedious labor. In the semi-automatic methods, the general procedure is obtaining terms, acquiring relations between terms, and thus the ontology can be constructed from all kinds of sources. The common techniques adopt heuristic rules which only acquire limited relations [Hearst 1992, Maedche 2000].

FCA (formal concept analysis) is a mathematical approach to data analysis based on the lattice theory. Because formal concept lattices are a natural representation of hierarchies and classifications, the orientation of FCA has turned from a pure mathematical tool towards computer science in the last few years [Stumme 2002], especially for automatic construction of ontologies [Cimiano 2004]. In this paper, we propose to construct a domain-specific ontology with FCA. How to apply FCA to the Chinese resources is the focus of this paper. To achieve this goal we have designed two experiments using two data resources.

The rest of the paper is organized as follows. Section 2 introduces the basic concepts and some related work. Section 3, describes the dataset selection and design of experiments. Section 4 presents the experimental results and discusses the analysis. Section 5 gives the concluding remarks and future directions.

2 Basic Concepts and Related Work

2.1 Ontology Definitions

In [Gruber 1995], an ontology was defined as "a specification of a conceptualization". The difference between ontology and conceptualization is that ontology is language-dependent while conceptualization is language-independent. According to Gruber's definition, the subject of *ontology* is the study of the *categories* of things that exist or may exist in some domain. Furthermore, an *ontology* is a catalog of the types of things that are assumed to exist in a domain of interest D from the perspective of a person who uses a language L for the purpose of talking about D. In the simplest case, an ontology describes a hierarchy of concepts related by subsumption relationships in some domain [Guarino 1998].

In [Sowa 2000], distinction is made between a formal ontology and an informal ontology. An informal ontology may be specified by a catalog of types that are either undefined or defined only by statements in a natural language. 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. In this paper, *ontology,* is defined as follows.

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. Thus the ontology in this paper refers to a formal ontology.

Normally, in ontology construction, both L and D are implicit because any construction method would be applied to a specific language, L, in a specific domain, D. Ontology construction methods aim at how to obtain C and how to build R. Of course, some of the ontology construction methods can be dependent on L, D, or both.

2.2 FCA Overview

FCA is a formal technique for data analysis and knowledge representation. It can be used to automatically construct formal concepts as a lattice for a given context, to replace the time-consuming manual building of domain ontology. FCA takes two sets of data, one is called the *object set* and the other is called the *attribute set*, to find a binary relationship between the data of the two sets, and further constructs a so-called *formal concept* lattice with a concept inclusion ordering according to a *formal context*. The definitions of *formal context* and *formal concept* in FCA [Ganter 1999] are defined as follows.

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 $A' := \{m \in M \mid (g,\mathrm{m}) \in I, \forall\, g \in A\}$ and $B' := \{g \in G \mid (g,\mathrm{m}) \in I, \forall m \in B\}$.

For a formal concept (A,B), A is called the *extent* and B the *intent* of the formal concept. Formal concepts satisfy the ***partial ordering relationship***, denoted by $\leq$, with regard to inclusion of their extents or inverse inclusion of their intents, formalized by:

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

It can be seen that a formal concept (A_1, B_1) in the concept lattice contains more attributes than its superconcept (A_2, B_2). On the other hand, more attributes a formal concept is associated with, fewer objects would belong to it. Thus, a formal concept in the lattice is associated with fewer objects and more attributes than a superconcept.

The whole formal concept lattice satisfies the partial ordering relations. When applying FCA to ontology construction, each term used in a specific domain can be mapped into an object in FCA. Thus, a term along with its set of attributes forms a node as a formal concept in the FCA lattice. Along the partial ordering relationships built based on the definition given earlier, relationships among different terms can be found. To see the mapping of FCA model to the definition of ontology, it can be seen that a formal concept in FCA corresponds to a concept in concept set C, and the partial ordering relationship in FCA corresponds to R, for a specific language L and a specific domain D. How to make use of the partial-ordering relation to get relations between terms is the subject of an application making use of the FCA model.

2.3 Related Work

FCA is an effective technique for construction of formal ontologies. In general, most work focus on the selection of formal objects and attributes. In [Haav 2003], a text describing a certain entity is seen as an object and thus an object used in FCA can be any domain-specific text that uses domain-specific vocabulary and describes domain-specific

entities. Attributes of an object are noun-phrases that are present in the domain-specific text. An ontology used in the real estate domain was then constructed. [Jiang 2003] explored the potential role of formal concept analysis (FCA) in a context-based ontology construction in a clinical domain. The medical documents are used to represent formal objects and the compound medical phrases extracted from NLP module are used to represent formal attributes. The result showed that 57.7% of the medical concept relations extracted were identified positive. [Quan 2004] proposes the FOGA (Fuzzy Ontology Generation frAmework) to incorporate fuzzy logic into FCA to form a fuzzy concept lattice for an academic semantic web. Documents and research topics (terms) are used as formal objects and attributes respectively. The relationship between an object and an attribute is no longer a binary value, but a membership value between 0 and 1. The works above all assume that one document focuses on one topic. Then the terms extracted from that document can be seen as the attributes of the topic. However, often one document has more than one topic, thus the result of ontology generation is not very good.

Some other researches take words or (terms) as objects. In [Cimiano 2003], verb-object dependencies are extracted from texts where the head word of objects are considered as FCA objects and the corresponding verbs together with the postfix "able" are used as attributes. [Priss 2004] gives a good summary of linguistic application of FCA and generalizes that lexical databases can often be represented or analyzed using FCA.

3 Dataset Selection and Design of Experiments

In order to construct an ontology with FCA, terms associated with certain concepts are represented by a set of formal objects. After correct selection of attributes, the relations between terms or concepts in the domain can then be extracted through the FCA lattice. Therefore, in this application, the most important issue is the selection of an appropriate attribute set, which is determined by data sources.

Generally speaking, data sources can either be lexicon base or corpus. When using the lexical base, each term is considered given in the lexical source and it is relatively easy to get an existed attribute set compiled by experts. Whereas using corpus-based approach, attributes are selected based on real data independent of possible bias which can be introduced because of human intervention. However, the choice of corpus can affect the result. Thus, a large-scale and domain representative corpus must be available in this method.

3.1 Dataset Selection

In this paper, the objective is to try FCA in constructing ontology from different data sources. The chosen domain is Information Technology with a given set of IT terms as the concept terms in the domain, denoted as T. This hand-picked set of domain specific terms T, with a total of 49 terms (see Appendix), are then used as formal objects in FCA. The main work is to explore the selection of attributes using different data sources for ontology construction. For comparison, one lexical base and one corpus are used in two separate experiments to examine the selections of attributes. It should be pointed out that experiments can only be conducted based on data sources available.

In the experiment using lexical source, HowNet [Dong 2000] is chosen as the data source. The HowNet lexicon can be seen as a Chinese lexical database and each entry in it represents a concept and is defined by related sememes, which are the smallest indivisible semantic units. For example, "part" is a sememe and "显示器(monitor)" is a term [1] defined by a set of sememes such as "computer", "part", "look". The HowNet version 2000, which is used in the experiments, includes about 1,667 sememes and 68,630 concepts. Although the subsumption relations between sememes have been given, there is no visualized view of concepts.

In the corpus-based experiment, a large-scale corpus is used as data source for ontology construction. The corpus is composed of several newspapers, segmented and tagged [Yu 2001], consisting of 97 millions words. The selection of attributes is purely statistical based on that words that co-occur with the IT terms with certain

[1] Here the hypothesis is that a term only has one meaning in a specific domain. Hereafter, a term is equal to a concept, which is different from a formal concept and only represents an object in FCA.

significance are chosen as the attributes. Thus word bi-gram co-occurrence database is established by collecting and sorting all word bi-gram co-occurrences within [-5, 5] context windows in the corpus, to support the construction of attribute sets.

Here, both HowNet and the large-scale corpus are not good IT domain-specific data sources, however, the goal of this work is not to construct a full IT ontology. Rather, the focus is on a given set of terms/concepts, how different data sources can affect the construction of ontology. Thus it is reasonable to choose two general purpose data sources. Besides, the terms selected are only the most representatives of the IT domain and they are also used often in general context and thus contained in the chosen data sources.

As the 49 terms were chosen already to serve as the concept terms in the IT domain, the focus now is converted to the selection of an attribute set. With different data sources, different attribute sets are used to construct the formal concept contexts. The Java API of an open source software ConExp[2] (Concept Explorer) is used to generate the concept oriented views for the data.

3.2 Experiment 1: Data Source from HowNet, with Sememes as Attributes

HowNet lexicon can be seen as a typical lexical base. Each term in its lexicon has a set of descriptive sememes, which can be used to define and discriminate the terms. Because HowNet is about general knowledge, many terms are of no interest for our work. Thus, only the IT-relevant terms were selected. In this experiment these IT terms are mapped into objects in FCA and sememes as attributes. The relationship between an object and an attribute is represented by a binary membership value. If an IT term t_i is defined by a sememe s_j, then the membership value $\mu(t_i, s_j)$ is 1, or 0 otherwise. Here is an example of the formal context illustrated in the form of a matrix as shown in the upper part of Figure 1 with a selected subset of terms from T. We use the reduced set of terms only to make the illustration more readable. In the figure, the rectangle with symbol "×" means the membership value of 1. The corresponding concept lattice is depicted with the tool ConExp as shown in the lower part of Figure 1. The lattice displays the extension of formal concepts represented by IT terms and the partial ordering relationships between formal concepts.

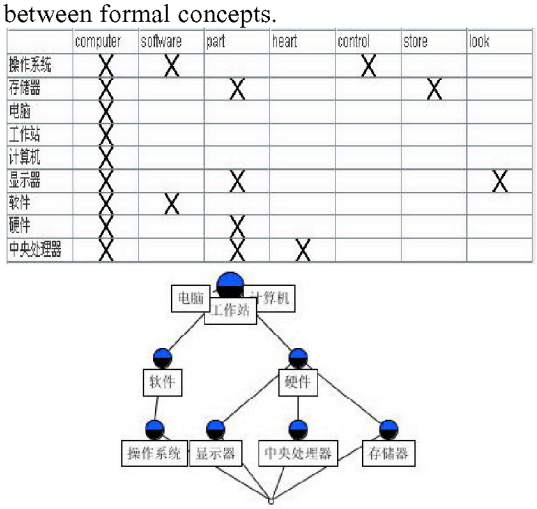

Figure 1. Example of concept context and concept lattice (terms as objects and sememes as attributes)

To further interpret the relationships of any two terms, two kinds of relations, superclass and equivalence, are defined. The objects with fewer attributes are considered as superclasses with respect to objects with more attributes. For example, "电脑(computer)" is the superclass of "硬件(hardware)" because the attribute set of "电脑(computer)" has only one attribute ("computer") which is contained in the attribute set ("computer", "part"), of "硬件(hardware)". Again, "硬件(hardware)" is the superclass of "硬盘(hard disk)" because the attribute set of "硬件(hardware)", ("computer", "part"), is contained in ("computer", "part", "store") which is the attribute set of "硬盘(hard disk)". Two objects described by the same set of attributes are considered as equivalent. For example, "电脑(computer)" and "计算机(computer)" are considered equivalent because they are described by the same set of attributes, ("computer"). Then, the relations are represented in the form of a triplet $<t_i, t_j, R(t_i, t_j)>$, where t_i and t_j are any two terms and $R(t_i, t_j)$ represents the relation, which can either be superclass or equivalence, The equivalence and superclass relations are not explicitly indicated in Figure 1. But, we would be able to find them and list them out according to the lattice result.

[2] http://sourceforge.net/projects/conexp

3.3 Experiment 2: Data Source from Corpus Using Context Verbs as Attributes

Due to the absence of IT-domain corpus, we use a processed huge-scale general corpus to simulate a domain-specific corpus because IT terms also often occur in general texts with their own specific meanings as explained in Section 3.1. For each term, the words in its context always have close relationships with it. Through linguistic observation, it is easy to see that usually content words play a very important role in describing a term semantically. Content words are determined by these POS tagging information which is available in the selected corpus. In principle, both verbs and nouns are content words and can be used as attributes. In this paper, a simple context window is used to observe context without regards to syntax or semantics. Because all the selected IT terms are considered nouns and are thus identified by the POS noun tag in the corpus, and if nouns are also chosen as attributes, it is very likely that the attribute set can contain terms, which can form anfractuous relations, making it difficult to acquire explicit hierarchical relations. Thus, the following experiment, only verbs in the context of a term is used as attributes.

As only verbs are considered in the context of IT terms as attributes, which verbs should be included in the attribute set of a term becomes the main issue. The approach is to make use of the co-occurrence statistics between terms and verbs. The co-occurrence statistics is represented by a triplet $<t_i, v_j, n_{ij}>$, where t_i represents a term in T, v_j is a co-occurring verb, and n_{ij} indicates their co-occurring frequency. After the collection of statistics, all the triplets are sorted in descending order according to n_{ij}. The system has a threshold parameter N below which the co-occurrence is considered statistically insignificant. Now the selection can begin, the algorithm picks all $<t_i, v_j, n_{ij}>$ with $n_{ij} \geq N$. However, if at this time, some term t_j in T still does not occur in the selected set, the algorithm goes further down the sorted list until all the terms in T has occurred at least once. Then all the verbs in the selected set are included in the attribute set. In order to avoid using some very general verbs as attributes, such as "是(is)", "成为(become)" which has no discriminating power in any specific domain unless further syntactic analysis is conducted, a stop word list is used to eliminate them in the attribute set.

The relationship between a term and a verb is also represented by a binary membership value. For every triplet $<t_i, v_j, n_{ij}>$ in the selected set, the membership value $\mu(t_i, v_j)$ is 1, 0 otherwise. Figure 2 illustrates the examples of concept context and concept lattice using a subset of T. In this figure, some of the attributes which does not affect the construction of concept lattice, referred to as reducible attributes [Ganter 1999] are already eliminated by ConExp, the tool used to draft the lattice.

Although the formal concepts satisfy the relationship of partial ordering, the subsumption relation between objects in Figure 2 has a direction reverse to that in Figure 1 to make it easier to read visually. Generally speaking, the more general meaning a term represents, the more co-occurring verbs it would have. Thus, in Figure 2, the nodes in the lower levels are actually superclasses of those in the upper levels.

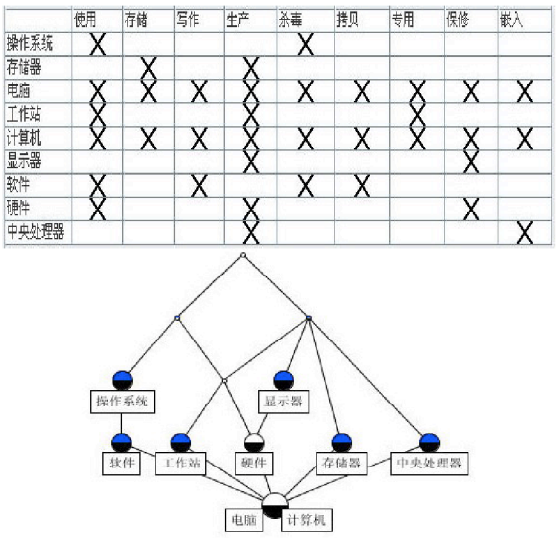

Figure 2. Example of concept context and concept lattice (terms as objects and context verbs as attributes)

4 Evaluation and Discussion

4.1 Evaluation

In Experiment 1, 33 attributes are extracted based on the descriptive sememes for those 49 terms in T from HowNet. In Experiment 2, with the same set

49 terms in *T*, the co-occuring verbs within [-5, 5] context windows are extracted as discussed in Section 3.3. The co-occurrence threshold *N* is set as 10 and 1,094 co-occurrence pairs are extracted from the corpus. Among them there are 326 distinct verbs. After filtering with the stop word list and removal of reducible attributes, only 68 verbs are left to serve as attributes. The two concept lattices generated respectively in two experiments are illustrated in Figure 3a and Figure 3b, respectively.

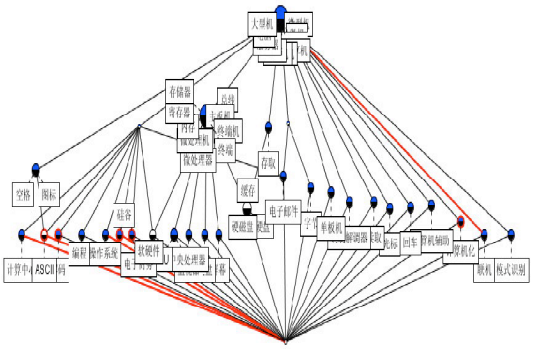

Figure 3a. Lattice generated in Experiment 1

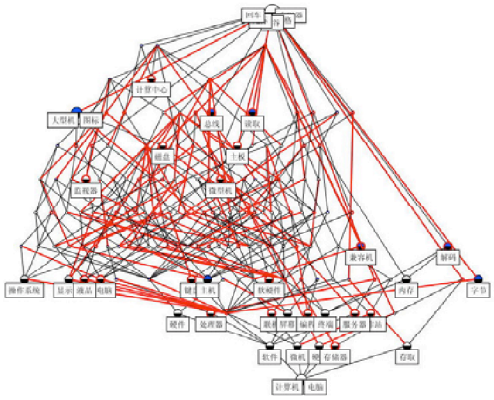

Figure 3b. Lattice generated in Experiment 2

Currently there is no a uniform evaluation measure on the ontology construction. Most researchers evaluate their ontologies in two parts including lexical part and taxonomic part [Cimiano 2003, Jiang 2003]. In [Cimiano 2003], he compares its ontology with standard domain ontology. The ontologies are seen as a semiotic sign system and present a comparison based on lexical overlap (LO) as well as taxonomic similarity between ontologies (SC). In [Jiang 2003], medical phrases and attribute implication pairs are respectively evaluated manually by the ratio of answers.

In this paper, it is assumed that the domain-specific terms are given and are considered correct. Thus the evaluation work mainly focuses on measuring the automatically generated taxonomic relationship between terms. According to the link in each concept lattice, the equivalence and superclass (the reverse is called subclass) relations between terms which are represented by a triplet $<t_i, t_j, R(t_i, t_j)>$. In Experiment 1, 118 of such relationship triplets are found. In Experiment 2, 73 of such triplets are found. Five researchers in the IT fields were asked to evaluate the eligibility of these triplets manually. For each relation represented by a triplet, an evaluator only needs to answer "YES" to mean that it is a correct relation, or "NO" otherwise. The evaluation results show that Experiment 1 gets about 43.2% of answers as "YES" and Experiment 2 gets about 56.2% of answer as "YES".

4.2 Discussion

Automatically identifying the relations between domain-specific concepts is the main task in ontology construction. Due to the lack of other data, comparisons can only be done between the two sets of results in the two experiments discussed in this paper. The analysis of the results is discussed here.

In principle, each attribute in the attribute set of an ontology should be independent. However, in reality, the selection of independent attributes is very difficult. For example, in HowNet, both "computer(计算机)" and "software(软件)" are sememes. But we all know semantically, they are related. In the selected attributes in Experiment 2, "设计(design)" and "开发(development)" can be another example of related attributes.

As mentioned earlier, 33 and 68 attributes are selected respectively in Experiment 1 and Experiment 2. Comparing the two lattices in Figure 3, it is obvious that Experiment 1 produces a relatively flat concept lattice with much less number of levels in terms of its hierarchical structure and less intricacy. This can be explained by saying that the granularity of the sememes are not fine enough to have enough discriminating power for different concept terms. Consequently, at the same level of hierarchy, more terms would

be equivalent. In fact, Experiment 1 has 86 equivalent classes whereas Experiment 2 has only 12 equivalent classes. For example, in HowNet, both "终端(terminal)" and "存储器(memory)" are described by the same attribute set (part, computer), and are thus equivalent. With a finer granularity, "终端(terminal)" should be defined by the sememes (part, computer, display), whereas, "存储器(memory)" should be defined by (part, computer, store). On the other hand, the number of superclass relationships in Experiment 1 is 32, 29 less than that in Experiment 2. Therefore, in addition to the number of attributes used in building an ontology, both the number of equivalence relationships and superclass relationships are also good measurement on the discriminating powers of attributes selected. The FCA model provides a very good visual method to observe the structure of these partial ordering relationship to examine whether an ontology is properly built. Furthermore, Experiment 1 has selected the manual HowNet lexicon which has cost a large amount of labor and it is not tailored for the IT domain. To modify it for the construction of a domain specific ontology, there would be a lot of manual work. In Experiment 2, a corpus is used as the data source. The corpus needs to be segmented and POS tagged beforehand. Because the techniques of segmentation and POS tagging are relatively mature, the whole process is basically automatic and thus saves a lot of manual work. Also, if the corpus is domain relevant and reasonable in size, the result should be quite reliable. In addition, subjectivity can be reduced to minimum when a large-scale domain relevant data is available.

For both the manually constructed ontology using HowNet and the automatically generated one based on a corpus, the ontology built still need manual tuning, because some of the data are certainly wrong from a semantic view. For example, in Experiment 1, since the term "存储器(memory)" is described by (part, computer) and the term "硬盘(hard disk)" is described by (part, computer, store), thus, naturally, it is wrong that memory would be the superclass of hard disk. Experiment 2 also has similar problems.

The ratio of correct relations (answer "YES") in Experiment 2 is higher than that of Experiment 1. However, this cannot be interpreted blindly to that manually constructed ontology is not as good as statistically obtained ontology. By looking at the different ways of constructing ontology, it can be seen that the more important issue in manual construction is the granularity of the attributes and the correct mapping of these attributes to the terms. On the other hand, it is more important to determine the types of attributes to be included for consideration. By looking at an example in Experiment 2, where term "微机(micro computer)" is described by (develop|开发, apply|应用, control|控制, design|设计), one would question how useful these attributes are, even if its discriminating power is higher than that in Experiment 1. If we change these attributes from verbs to the corresponding nouns, (development, application, control, design), this set only represents the predicate nature of the term. Relationships, such as part-of, are not necessarily present. This gives rise to an issue of algorithm design of automatic ontology extraction. Basically, clustering verbs only is not good enough as it does not semantically make very good sense. The use of nouns and noun phrases are inevitable. Furthermore, more complex algorithms should look into the context of the terms both syntactically and semantically so that subject and object relationships can be identified.

5 Conclusions and Future Work

In this paper, the FCA model is used to help the analysis of ontology constructed using different data sources. Two experiments are made to gain insight as to what are the important issues in ontology construction and trade-off in efficiency. It can be seen that in manual ontology construction, both the choice of sememes and their granularity are important as well as the correct mapping of the terms to the set of attributes. On the other hand, the methods to select the types and related context words in the corpus are important. Pure statistical method without regards to syntax and semantics would result in an ontology which is difficult to interpret. It is also shown that the use of visual tool is very helpful in building a good ontology as the superclass and equivalence relationships can help to tune the ontology to be built.

There are two future directions of work. Firstly, more comprehensive algorithms will be investigated using more syntactic and semantic cues to improve the quality of ontology

construction. With the temporal information to be taken into consideration, algorithms on terms extraction and their mapping into an existing ontology can also be developed. Secondly, the use and the enhancement of the visualizations tools will also be exploited to improve the quality of ontology construction and the extraction of useful relationships.

Acknowledgement

The work presented in this paper is supported by Research Grants Council of Hong Kong (reference number: CERG PolyU 5190/04E and 5181/03E) and the National 863 Programme (Grant no: 2001AA114210).

References

Bain, M., 2003, Inductive Construction of Ontologies from Formal Concept Analysis, In: T. Gedeon and L. Fung, editors, AI 2003: Proc. of the 16th Australian Joint Conference on Artificial Intelligence, LNAI 2903, pages 88--99, Berlin. Springer.

Cimiano, P. & Staab, S. & Tane, J., 2003, Automatic Acquisition of Taxonomies from Text: FCA Meets NLP. In Proceedings of the International Workshop on Adaptive Text Extraction and Mining.

Cimiano, P. & Hotho, A. & Stumme, G. & Tane. J., 2004, Conceptual Knowledge Processing with Formal Concept Analysis and Ontologies. In Proceedings of the 2nd International Conference on Formal Concept Analysis.

Dong, Z. Dong, Q. *HowNet,* http://www.keenage.com

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

Gruber, T.R. 1995, Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human and Computer Studies, 43(5/6):907-928.

Guarino, N. (ed.), 1998, Formal Ontology in Information Systems. Proceedings of FOIS-98, Amsterdam: IOS Press.

Haav, H-M., 2003, An Application of Inductive Concept Analysis to Construction of Domain-specific Ontologies, In: B. Thalheim, Gunar Fiedler (Eds), Emerging Database Research in East Europe, Proceedings of the Pre-conference Workshop of VLDB 2003, Computer Science Reports, Brandenburg University of Technology at Cottbus, 14/3, pp 63-67.

Hearst, M.A., 1992, Automatic Acquisition of Hyponyms from Large Text Corpora. In Proceedings of the 14th International Conference on Computational Linguistics.

Jiang, G. & Ogasawara, K. & et al., 2003, Context-based ontology building support for clinical domains using formal concept analysis. Int J Med Inform. 71(1):71-81.

Maedche, A. & Staab, S., 2000, Discovering Conceptual Relations from Text. ECAI 2000. Proceedings of the 14th European Conference on Artificial Intelligence, IOS Press, Amsterdam.

Priss, U. 2004, Linguistic Applications of Formal Concept Analysis. In G. Stumme and R. Wille, editors, Formal Concept Analysis - State of the Art. Springer.

Quan, T.T. & Hui, S.C. & Cao, T.H., 2004, FOGA: A Fuzzy Ontology Generation Framework for Scholarly Semantic Web, Knowledge Discovery and Ontologies (KDO-2004), Workshop at ECML/PKDD 2004.

Sowa, J.F., 2000, Knowledge Representation, Logical, Philosophical, and Computational Foundations, Brooks/Cole Thomson Learning.

Stumme, G., 2002, Formal Concept Analysis on its Way from Mathematics to Computer Science. In: U. Priss, D. Corbett, G. Angelova (Eds.): Conceptual Structures: Integration and Interfaces, Proc. ICCS 2002, LNAI 2393, Springer, Heidelberg 2002, 2-19.

Yu, S.W et al., 2001,Guideline of People's Daily Corpus Annotation, Technical report, Beijing University, 2001 (in Chinese)

Appendix: IT terms selected

ASCII	CPU	编程 (programming)
操作系统 (operating system)	存储器(memory)	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)
显示器(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(hardward 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)		