Ontology Construction Based on Latent Topic Extraction in a Digital Library

Jian-hua Yeh and Naomi Yang

Department of Computer Science and Information Engineering
Aletheia University
au4290@email.au.edu.tw
Graduate Institute of Library and Information Studies
National Taiwan Normal University
shupiny@ms2.hinet.net

Abstract. This paper discusses the automatic ontology construction process in a digital library. Traditional automatic ontology construction uses hierarchical clustering to group similar terms, and the result hierarchy is usually not satisfactory for human's recognition. Human-provided knowledge network presents strong semantic features, but this generation process is both labor-intensive and inconsistent under large scale scenario. The method proposed in this paper combines the statistical correction and latent topic extraction of textual data in a digital library, which produces a semantic-oriented and OWL-based ontology. The experimental document collection used here is the Chinese Recorder, which served as a link between the various missions that were part of the rise and heyday of the Western effort to Christianize the Far East. The ontology construction process is described and a final ontology in OWL format is shown in our result.

1 Introduction

Manual ontology construction has been a labor-intensive work for human beings, since humans are capable of creating semantic hierarchy efficiently. But with the growing size of real world concepts and their relationships, it is more difficult for humans to generate and maintain large scale ontologies. Meanwhile, the quality of the knowledge structure in an ontology is hard to maintain because human is not able to keep the criteria of concept creation consistently. Therefore, human-generated knowledge networks are usually difficult to span, such as web directories, large organization category hierarchies, and so forth. Our experiences on constructing web ontology [1] and government ontology [2] also show that it is difficult to generate fully semantic-aware concept hierarchy purely replying on traditional data clustering algorithms. In this paper, we introduce an effective process to construct domain ontology automatically based on a special collection called the Chinese Recorder [5], which served as a link between the various missions that were part of the rise and heyday of the Western effort to Christianize the Far East. A special ontology construction process is designed and a final ontology described in OWL [3] format is shown in our result. This paper aims at two major issues, as listed below:

G. Buchanan, M. Masoodian, S.J. Cunningham (Eds.): ICADL 2008, LNCS 5362, pp. 93–103, 2008.

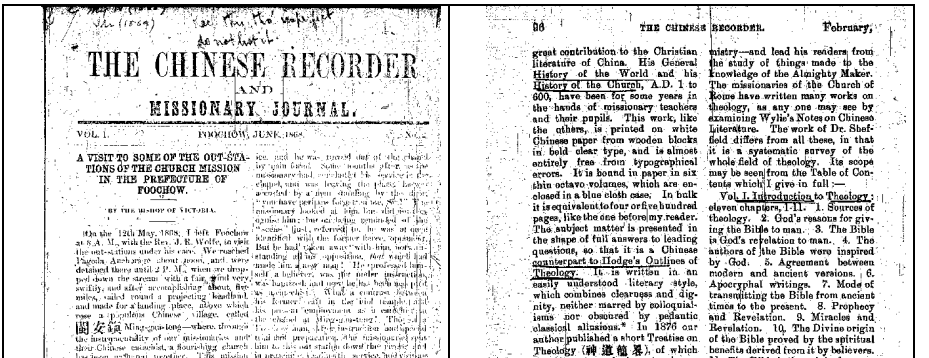
1. Create an effective process for ontology construction of historical documents Generally speaking , there are two ways to generate ontologies: manual and automatic construction. In recent years, a number of related discussions focus on the process of manual generation of ontologies, one of the most frequently referenced article is “Ontology 101” [4]. In [4], a complete ontology creation process is introduced, with the standard steps including determination of the domain and scope, consideration of reuse, enumeration of important terms, definition of the classes and the class hierarchy, definition of the slots, and creating instances. For the process of automatic construction of ontologies, many algorithms were proposed and developed, which will be discussed in the next section. One of the major goals in our research aims at aged historical collections, trying to develop an effective and automatic process to construct domain ontology. This process will relieve the burden of domain expert with decreasing the time consumption and increasing the collection scale.
2. Construct knowledge network for historical collections
   The knowledge contained in historical documents is both rich and wide-ranging, which leads the research focus of creating knowledge network or knowledge hierarchy. But most of the researches produce content classification only, which is called taxonomy in that scenario. The taxonomy only represents the tree structure of content classification, which lacks of variety of relationships among concepts. That is, no complex knowledge structure contained in such kind of structure. One of the major goals in our research aims at creation of complex structure to represent rich knowledge in historical collections and description of domain ontology using W3C OWL standard.

2 Issues of Ontology Construction Process for the Chinese Recorder

Before going into the introduction of ontology construction process, the special historical collection in our research is described first. We use the collection called “The Chinese Recorder” for our domain ontology construction experiment. The Chinese Recorder is a valuable source for studying the missionary movement in China and the effect the missions had on shaping Western perceptions of and relations with the Far East. It was publis -language publication to last for 72 years. It served as a link between the varioushed monthly by the Protestant missionary community in China, and was the only English missions that were part of the rise and heyday of the Western effort to Christianize the Far East. It provided information about individual missionaries and mission activities, recounting their progress on evangelical, educational, medical, and social fronts. It featured articles on China's people, history, and culture. The Robert’s library of Toronto University has holdings for June 1868-December 1876 and Jan 1915- December 1932, while the Chinese Church Research Center (Taiwan) has holdings for all the collections among 72 years.

Currently, most of the Chinese Recorder collection copies are kept in microfilm form. In the early years of its publication (1860s), the printing methods is pretty primitive and we found that the scanning quality of the microfilms is poor (please refer to Fig. 1). If the optical character recognition (OCR) applied on such digital images,

many kinds of noise (mostly from the dirty spots on the pages) will affect the recognition effectiveness. Besides, the page formats of the Chinese Recorder varies from volume to volume (single-column, double-column, etc.), which creates certain barrier of whole chapter text extraction. This problem is also an important issue in related researches. Since the process proposed in this paper concerns only the relationship between document and terms contained in the document, the whole chapter text extraction is unnecessary and the noise effect can be controlled in our experiment.



THE CHINESE RECORDER
AND
MISSIONARY JOURNAL.

VOL. I. FOOCHOW, JUNE, 1868. No. 2.

A VISIT TO SOME OF THE OUT-STATIONS OF THE CHURCH MISSION IN THE PREFECTURE OF FOOCHOW.

BY THE RT. REV. THE BISHOP OF VICTORIA.

On the 12th May, 1868, I left Foochow at 8 A. M., with the Rev. A. R. Wolfe, in order to visit some of the out-stations under his care. We reached Pagoda Anchorage about noon, and were detained there until 2 P. M., when we dropped down the stream with a fair wind very swiftly, and after accomplishing about five miles, sailed round a projecting headland and made for a landing place, above which rose a populous Chinese village, called 閩安鎮 Mingan-tsing—where through

THE CHINESE RECORDER. February,

great contribution to the Christian literature of China. His General History of the World and his History of the Church, A.D. 1 to 600, have been for some years in the hands of missionary teachers and their pupils. This work, like the others, is printed on white Chinese paper from wooden blocks in bold clear type, and is almost entirely free from typographical errors. It is bound in paper in six thin octavo volumes, which are enclosed in a blue cloth case. In bulk it is equivalent to four or five hundred pages, like the one before my reader. The subject matter is presented in the shape of full answers to leading questions, so that it is a Chinese counterpart to Hodge's Outlines of Theology. It is written in an easily understood literary style, which combines clearness and dignity, neither marred by colloquialisms nor obscured by pedantic classical allusions." In 1876 our author published a short Treatise on Theology (神道總基), of which the present work is an enlargement. It is the autumnal fruit of study and experience continued through a score of years. The author, instead of making a translation of some foreign work, has prepared a compendious treatise adapted to the wants and comprehension of Chinese theological students. Dr. Martin, of Peking, in 天道溯原, Heavenly Doctrine traced to its Source, and Dr. John, of Hankow, in 進窄入門, The Gate of Virtue and Wisdom, have prepared two excellent works on the Evidences of Christianity. Dr. Alexander Williamson, lately deceased, has in 格物探原 given us a valuable treatise on Natural Theology. His plan is to take a wide discursive over a given field—astronomy, geology, botany or che-

*As to terms, 上主 and 主 are used for *God*; 聖靈 for *Holy Spirit*; 斯世 for *Humanity* as opposed to 人性 *Humanity*.

istry—and lead his readers from the study of things made to the knowledge of the Almighty Maker. The missionaries of the Church of Rome have written many works on theology, as any one may see by examining Wylie's Notes on Chinese literature. The work of Dr. Sheffield differs from all these, in that it is a systematic survey of the whole field of theology. Its scope may be seen from the Table of Contents which I give in full :—

Vol. I. Introduction to Theology : eleven chapters, 1-11. 1. Sources of theology. 2. God's reasons for giving the Bible to man. 3. The Bible is God's revelation to man. 4. The authors of the Bible were inspired by God. 5. Agreement between modern and ancient versions. 6. Apocryphal writings. 7. Mode of transmitting the Bible from ancient times to the present. 8. Prophecy and Revelation. 9. Miracles and Revelation. 10. The Divine origin of the Bible proved by the spiritual benefits derived from it by believers. 11. The Bible is the original source of the great doctrine of man's salvation.

Vol. II. Theology : five chapters, 12-16. 12. Sources of man's knowledge of the Lord of the universe. 13. God's method of revealing Himself. 14. Errors concerning the nature of God. 15. The attributes of God. 16. The doctrine of the Trinity.

Vol. III. Cosmology : Eight chapters, 17-24. 17. God's creation of the universe. 18. Angels and evil spirits. 19. God's preservation and government of His creatures. 20. The decrees of God. 21. The creation of man. 22. Man's free will. 23. The sources of our judgments as to right and wrong. 24. The fall : literally, how man became drowned in sin.

Vol. IV. Soteriology : twelve chapters, 25-36. 25. The incarnation of Jesus and its relation to man. 26. The dignity of man re-

Fig. 1. The scanning quality of the Chinese Recorder microfilms is poor. (left: May 1868, right: February 1894).

When the text extraction finished, all the textual data in the Chinese Recorder are collected. The next issue will be the ontology creation process based on the extracted text. The automatic hierarchy construction researches have gained much attention in recent years [6,7]. Several classification and clustering methods have been proposed to solve the hierarchy generation problem [8,9,10,11]. Among these methods, the hierarchical agglomerative clustering (HAC) algorithms [7,12] attracts a lot of attentions since the clustering result is presented in the form of a tree structure, that is, a hierarchy is created. This kind of algorithms has great potential in automatic hierarchy generation since it keeps both scalability and consistency. While this is a promising technology, it still suffers certain shortage in the current development: the cluster generated by automatic process seldom keeps semantics, since the automatic process based on information retrieval (IR) techniques is usually term-based, not concept-based. This characteristic causes the clustering result seldom semantic-aware, which

is often not acceptable by human recognition. The automatic generation process can be improved when information retrieval and computational linguistics techniques advance, but for now, it is not applicable to generate high-quality knowledge networks through traditional clustering methods only. In recent years, the researches about topic extraction from texts are getting more and more attention, especially a promising technology called latent topic discovery. Latent topic discovery is invented to overcome the bottleneck of bag-of-words processing model in information retrieval area, trying to advance the text processing technology from pattern to semantic calculation.

For the researches in latent topic discovery, most of the research focuses aim at topic detection in text data by using term distribution calculation among the documents. Several important algorithms were developed, including Latent Semantic Analysis (LSA)[13], Probabilistic Latent Semantic Analysis (pLSA)[14], and Latent Dirichlet Allocation (LDA)[15]. LSA is one of the semantic analysis algorithms which differs from traditional term frequency-inverse document frequency (TF-IDF) model. The TF-IDF model consider the term frequency only, but the calculation of LSA combines some latent factor of textual data by adding additional vector space features such as singular value decomposition (SVD) of document-term matrix to analyze the document-term relationships. pLSA model is proposed to overcome the disadvantage found in by LSA model, trying to decrease the degree of computation by using probabilistic approach. pLSA analyzes the document-term relationships using latent topic space, just like LSA, which projects the term tj in set T together with document di in set D to a set of k latent topics Tk. pLSA and LSA try to represent the original document space with a lower dimension space called latent topic space. In Hofmann [14], P(Tk | d) is treated as the lower dimension representation of document space, for any unseen document or query, trying to find the maximum similarity with fixed P(t | Tk). Other than LSA and pLSA, the algorithm of Latent Dirichlet Allocation (LDA) is more advantageous since LDA performs even better than previous research results in latent topic detection. In fact, LDA is a general form of pLSA, the difference between LDA and pLSA model is that LDA regards the document probabilities as a term mixture model of latent topics. Girolamin and Kaban [16] shows that pLSA model is just a special case of LDA when Dirichlet distributions are of the same.

Because the latent topic discovery procedure is capable of finding semantic topics, we can further group these topics, regarding it as a semantic clustering process. All we need is to calculate the cosine similarity [17] between topics since the latent topics resolve the problems of synonym and polysemy, that is, the terms grouped under a latent topic reveal the term usage for some specific concept.

From the discussion above, it is necessary to design a latent topic based ontology construction process to ensure the successful ontology generation and overcome the difficulties created by old historical collections. In the next section, a latent topic extraction method is proposed to support automatic domain ontology construction.

3 The Proposed Method

This paper aims at developing an automatic domain ontology construction process for historical documents. We will also describe the final ontology with semantic web

related standards to show the knowledge structures. Before developing the construction process, all the Chinese Recorder microfilms are scanned and produce a large amount of digital images. The processing steps of this research are shown below:

1. Textual data generation
   We adopt standard OCR procedure in this step to generate raw textual data. As mentioned earlier, a large amount of digital images were generated, and we use OmniPage® as our OCR software to produce the raw text. We found that there is about 8%-10% errors in the raw text, so we design a statistical correction methodology to correct these errors via bigram correlation model [18], as described below:
   (a) A corpus [19] is chosen to get the term bigram data, a list of 65,000 entries is fetched and form a bigram matrix B. The value of each matrix cell B(termi, termj) is logp(the value p stands for the probability of termj after termi).
   (b) Match the raw text with the corpus term list fetched in (a), finding the unknown term U (the possible error word) and considering with previous/next terms F and R. List possible candidate words of U based on Levenshtein distance(a kind of edit distance)[23] (U1, U2, …, Uk) and calculate b = B(F, Ui)*B(Ui, R), the bigger value of b shows the more possible correction term. For the consecutive term errors, we do not process them since they are below the statistical threshold in our experiment(less than 0.5%).

   The second part of text generation is from the contents of The Chinese Recorder Index [20]. It contains three indexes: the Persons Index, the Missions and Organizations Index, and the Subject Index, which give call number, title, date, month and the year. Since this is a relatively young publication (published in 1986) with both good publication and preservation conditions, it is easy to fetch three kinds of term list via OCR process.

2. Latent topic extraction
   After the raw text is generated and corrected, we are ready for latent topic extraction. In this step, the Latent Dirichlet Allocation (LDA) is used to extract latent topics from raw text generated in previous step since LDA performs smoother topic range calculation then LSA and pLSA [16]. In this research, we treat every page in the Chinese Recorder as a basic data unit called “page document”. These page documents will form a document-term matrix known in information retrieval domain, generally a sparse matrix. Then the LDA estimation starts and the latent topics are generated.
3. Topic clustering
   In topic clustering step, we group the latent topics into higher level topics in a hierarchical manner. Because the latent topic contain semantics, so the clustering process is regarded as some kind of semantic clustering. In this research, the basic cosine similarity with hierarchical agglomerative clustering (HAC)[7,12] is adopted to generate high level topics called "super topics". The cosine similarity is calculated as follows:

Let $t_u = \{w_{u,1}, \cdots, w_{u,n}\}$ and $t_v = \{w_{v,1}, \cdots, w_{v,n}\}$ be two vectors of correlation values for the topic *u* and *v*, the topic similarity estimation function is

$$sim(u,v) = \frac{\vec{t}_u \cdot \vec{t}_v}{|\vec{t}_u| \times |\vec{t}_v|} = \frac{\sum_{i=1}^{n} w_{u,i} \times w_{v,i}}{\sqrt{\sum_{i=1}^{n} w_{u,i}^2} \times \sqrt{\sum_{i=1}^{n} w_{v,i}^2}}$$

The clustered super topics form a tree structure, but the whole ontology needs not to be a tree structure since the relationships among page documents, index terms, and latent topics are not simply hierarchical but graph-based, as shown in Fig. 2.

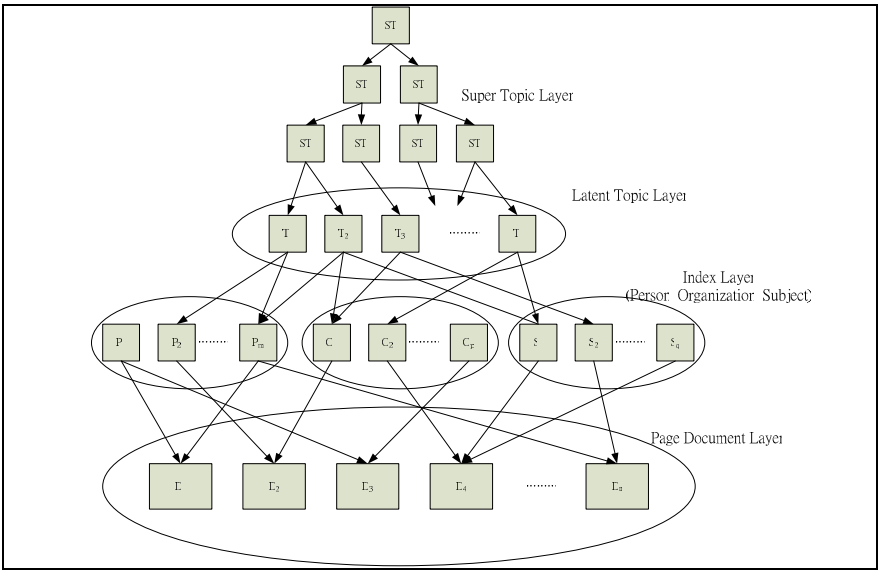


Fig. 2. Ontology layer and structure in this research

In Fig. 2, the nodes contained in page document layer, index layer, and latent topic layer form a graph structure. An additional feature shows that not all index terms will be referenced by latent topics because the latent topic choose terms based on obvious frequency of term occurrence.

4. OWL generation and domain expert revision

In this step, we define the OWL classes and properties for the domain ontology of the Chinese Recorder. According to the characteristics of data generated in previous steps, we propose the class and property definitions as below:

(a) *PageDocument* class: every page text fetched from the Chinese Recorder is an instance of this class.

(b) *Person* class: every person name shown in the Chinese Recorder is an instance of this class.
(c) *Organization* class: every organization or mission name shown in the Chinese Recorder is an instance of this class.
(d) *Subject* class: every subject shown in the Chinese Recorder is an instance of this class.
(e) *Topic* class: the extracted latent topics are the instances of this class. Besides, both topic and super topics are of the instances of this class.

For OWL property, we define:
(a) *Occurrence* property: this is the relationship between the index term (person, organization, and subject) and the page document with the index term.

The relationships between class/property/instance are shown in Fig. 3.

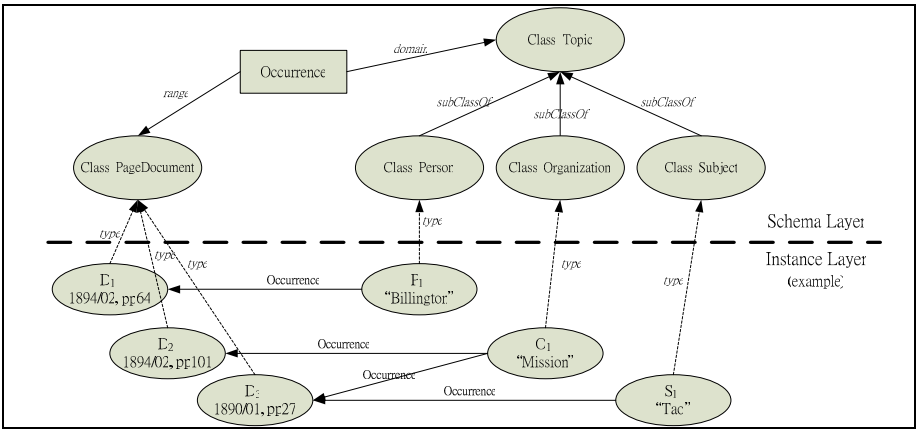


Fig. 3. Class/property/instance relationships in the Chinese Recorder ontology

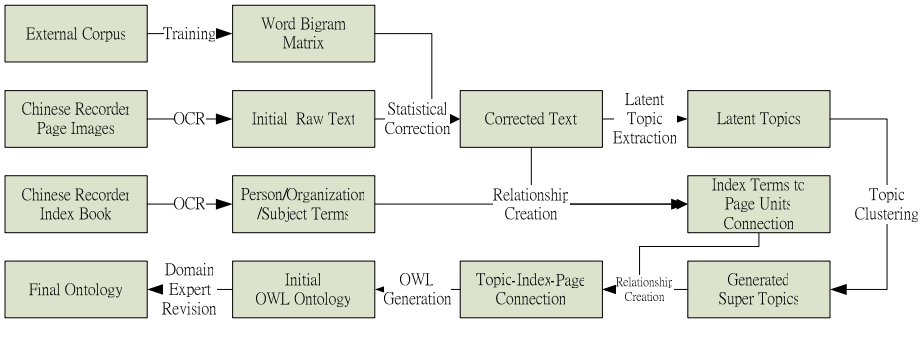


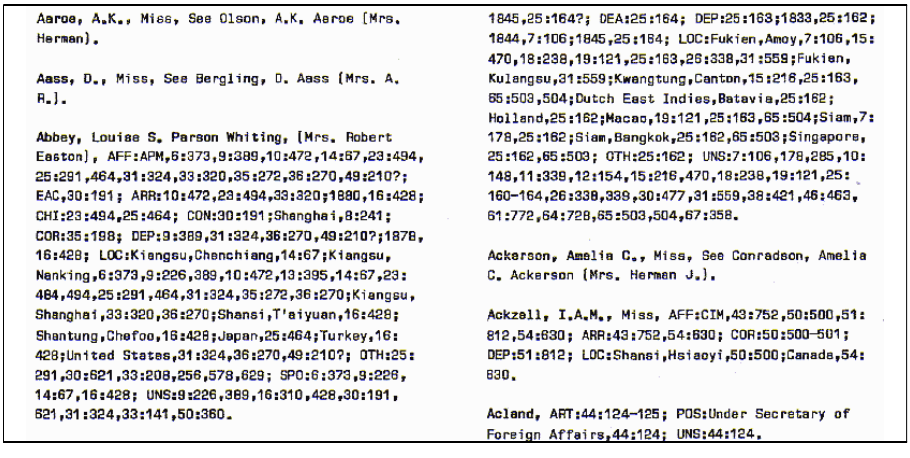
Fig. 4. The ontology processing steps in this research

With the above definitions in OWL, we can further organize the content generated in previous steps into a large graph structure, i.e. domain ontology of the Chinese Recorder. Because the data is OWL-based, it is possible to import into ontology editor such as Protégé[21] for further maintenance by domain experts.

The whole processing steps are shown in Fig. 4.

4 The Experiment

In this paper, we use a condensed dataset as a preliminary experiment to prove the correctness of our design. The 1890 and 1899 publications of the Chinese Recorder are chosen as our experimental data. This data set contains 204 page documents with several bad-condition pages (noises in pages). The statistical correction procedure mentioned in last section is adopted to increase the correctness of textual data. Besides, there are also non-English characters (such as Chinese contents, as shown in Fig. 1), but they are omitted in this experiment. For the index term generation, we found that the publication quality of the Chinese Recorder Index Book is pretty well, as shown in Fig. 5.



Aaroe, A.K., Miss, See Olson, A.K. Aaroe (Mrs. Herman).

Aass, D., Miss, See Bergling, D. Aass (Mrs. A. R.).

Abbey, Louise S. Parson Whiting, (Mrs. Robert Easton), AFF:APM,6:373,9:389,10:472,14:67,23:494, 25:291,464,31:324,33:320,35:272,36:270,49:210?; EAC,30:191; ARR:10:472,23:494,33:320;1880,16:428; CHI:23:494,25:464; CON:30:191;Shanghai,8:241; COR:35:198; DEP:9:389,31:324,36:270,49:210?;1878, 16:428; LOC:Kiangsu,Chenchiang,14:67;Kiangsu, Nanking,6:373,9:226,389,10:472,13:395,14:67,23: 484,494,25:291,464,31:324,35:272,36:270;Kiangsu, Shanghai,33:320,36:270;Shansi,T'aiyuan,16:428; Shantung,Chefoo,16:428;Japan,25:464;Turkey,16: 428;United States,31:324,36:270,49:210?; OTH:25: 291,30:621,33:208,256,578,629; SPO:6:373,9:226, 14:67,16:428; UNS:9:226,389,16:310,428,30:191, 621,31:324,33:141,50:360.

1845,25:164?; DEA:25:164; DEP:25:163;1833,25:162; 1844,7:106;1845,25:164; LOC:Fukien,Amoy,7:106,15: 470,18:238,19:121,25:163,26:338,31:559;Fukien, Kulangsu,31:559;Kwangtung,Canton,15:216,25:163, 65:503,504;Dutch East Indies,Batavia,25:162; Holland,25:162;Macao,19:121,25:163,65:504;Siam,7: 178,25:162;Siam,Bangkok,25:162,65:503;Singapore, 25:162,65:503; OTH:25:162; UNS:7:106,178,285,10: 148,11:339,12:154,15:216,470,18:238,19:121,25: 160-164,26:338,339,30:477,31:559,38:421,46:463, 61:772,64:728,65:503,504,67:358.

Ackerson, Amelia C., Miss, See Conradson, Amelia C. Ackerson (Mrs. Herman J.).

Ackzell, I.A.M., Miss, AFF:CIM,43:752,50:500,51: 812,54:630; ARR:43:752,54:630; COR:50:500-501; DEP:51:812; LOC:Shansi,Hsiaoyi,50:500;Canada,54: 630.

Acland, ART:44:124-125; POS:Under Secretary of Foreign Affairs,44:124; UNS:44:124.

Fig. 5. A partial page example of The Chinese Recorder Index

Next, the textual data generation step proposed in last section combined with statistical correction is adopted to generate 204 page documents with a total number of 95,005 words. For index term generation, a number of person, organization, and subject terms related to the document set are fetched. Related data statistics is shown in Table 1.

Table 1. Related data statistics in our experiment

Raw text	Person index	Organization index	Subject index
95,005	544	150	180

In latent topic extraction step, 80 latent topics are extracted from 204 example page documents. These topics form the basis of topic clustering step. Partial topic list is shown in Fig. 6.

Topic: Love Divine Kuling hours sacred God spiritually
Topic: children education school responsibility poor matter members
Topic: Christian Conference Spirit movement time ground Chinese
Topic: schools colleges prevent leading less American Christian
Topic: Bible preaching why unable expect comment Chinese
Topic: England historical bishop sufficient episcopal Church every
Topic: God human Go personal everywhere everything judgment
Topic: ancient Greeks contained seventh long fourth ear
Topic: worship days Yellow light According sect sun
Topic: method instance personal learn preaching employed first

Fig. 6. Partial latent topics generated by this experiment

In topic clustering step, we adopt cosine similarity as the decision function of hierarchical agglomerative clustering (HAC) algorithm to group the latent topics into "super topics", forming the final topic hierarchy (as shown in Fig. 7). According to the index-page-topic relationships described in previous section (as shown in Fig. 4), an OWL draft of the Chinese Recorder domain ontology is generated. This ontology draft is now ready to import into Protégé software for further maintenance (the screenshot is shown in Fig. 8).

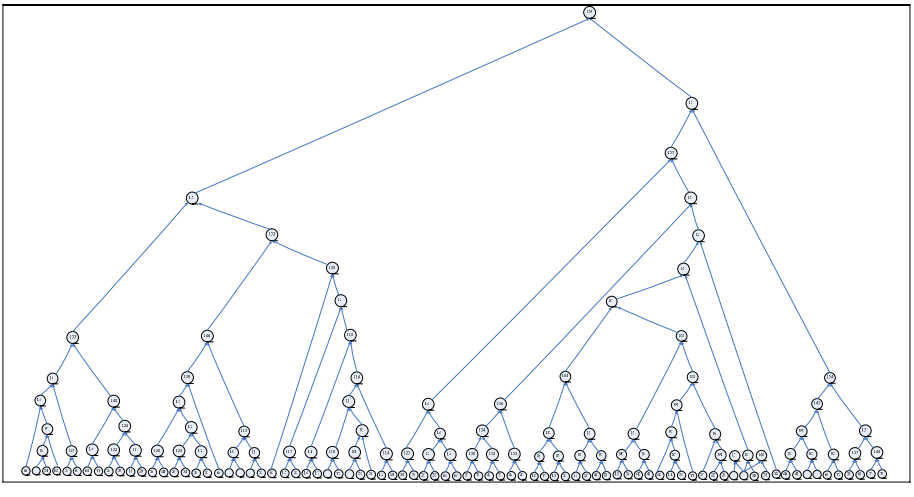


Fig. 7. Topic dendrogram in our experiment

After import into Protégé software, we are ready to give the ontology generation result to domain experts for further optimizations including revision and maintenance. The aspects of optimization include: 1. the original ontology classes contains Person, Organization, Subject, and Topic, the domain experts are able to refine the class hierarchy by adding more suitable classes such as deriving Event or Location subclasses from Topic class. 2. The derived classes form richer knowledge structure, which is able to add more semantics for later OWL instances.

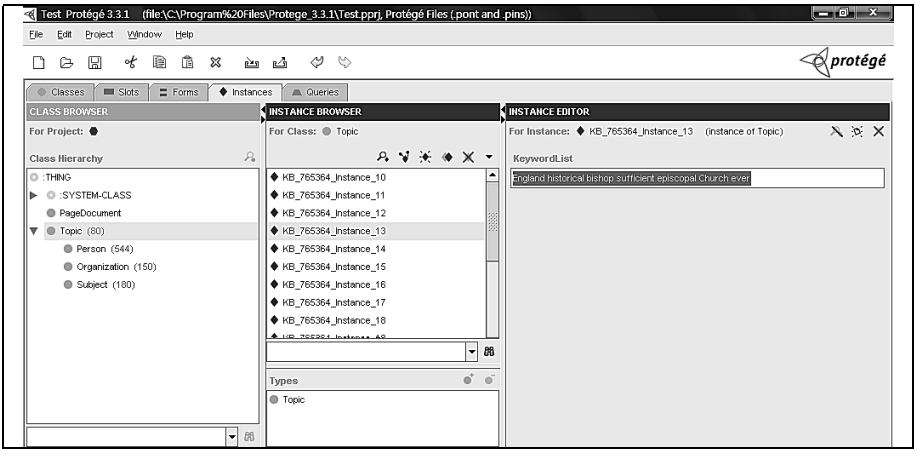


Fig. 8. OWL generated in this experiment can be imported into Protégé for further maintenance

5 Conclusion and Future Work

From the observation of the experimental result in this paper, we conclude that the construction of domain ontology draft is possible for historical collections through the introduction of proper information technologies. The extraction of semantics in documents shows the possibility of automatic ontology processing which minimize the human efforts in traditional ontology creation, while keeps semantics consistency with ontology generation. Meanwhile, the proposed processing steps introduce latent topic discovery as a basis of knowledge extraction, achieving both keeping important knowledge features of collection and relieving huge amount of human efforts.

The future works of this research contains: 1. expanding the data processing ranges for the whole Chinese Recorder collection, trying to create a complete domain ontology of the Chinese Recorder. 2. Describe the domain ontology with OWL-based format and published in our project web site to facilitate the content researches of the Chinese Recorder collection constantly. Besides, we will continue to improve our ontology processing algorithm to generate ontology with better semantic quality. For example, we are planning to adopt concept clustering [22] as the replacement of term vector similarity in topic clustering step through similarity propagation of term relationships.

References

1. Yeh, J.-H., Sie, S.-h.: Towards automatic concept hierarchy generation for specific knowledge network. In: Ali, M., Dapoigny, R. (eds.) IEA/AIE 2006. LNCS (LNAI), vol. 4031, pp. 982–989. Springer, Heidelberg (2006)
2. Chen, C.-c., Yeh, J.-H., Sie, S.-h.: Government ontology and thesaurus construction: A taiwanese experience. In: Fox, E.A., Neuhold, E.J., Premsmit, P., Wuwongse, V. (eds.) ICADL 2005, vol. 3815, pp. 263–272. Springer, Heidelberg (2005)

3. Deborah, L., McGuinness, Harmelen, F.v.: OWL Web Ontology Language Overview. W3C Recommendation (February 2004), http://www.w3.org/TR/owl-features/
4. Noy, N.F., McGuinness, D.L.: Ontology Development 101: A Guide to Creating Your First Ontology (2001)
5. The Chinese Recorder, Scholarly Resources, Inc, 1867-1941
6. Jain, A.K., Dubes, R.C.: Algorithms for clustering data. Prentice-Hall, Englewood Cliffs (1988)
7. Jain, A.K., Murty, M.N., Flynn, P.J.: Data Clustering: A Review. ACM Computing Surveys 31, 264–323 (1999)
8. Koller, D., Sahami, M.: Hierarchically classifying documents using very few words. In: Proceedings of ICML 1997, 14th International Conference on Machine Learning (1997)
9. Li, F., Yang, Y.: A loss function analysis for classification methods in text categorization. In: The Twentieth International Conference on Machine Learning (ICML 2003), pp. 472–479 (2003)
10. Valdes-Perez, R.E., et al.: Demonstration of Hierarchical Document Clustering of Digital Library Retrieval Results. In: Joint Conference on Digital Libraries (JDCL 2001), Roanoke, VA, June 24-28 (2001)(presented as a demonstration)
11. Yang, Y., Zhang, J., Kisiel, B.: A scalability analysis of classifiers in text categorization. In: ACM SIGIR 2003, pp. 96–103 (2003)
12. Widyantoro, D., Ioerger, T.R., Yen, J.: An Incremental Approach to Building a Cluster Hierarchy. In: Proceedings of the 2002 IEEE International Conference on Data Mining, ICDM 2002 (2002)
13. Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science 41(6), 391–407 (1990)
14. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42(1), 177–196 (2001)
15. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research 3(5), 993–1022 (2003)
16. Girolami, M., Kaban, A.: On an equivalence between PLSI and LDA. In: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 433–434 (2003)
17. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, Reading (1999)
18. Collins, M.: A new statistical parser based on bigram lexical dependencies. In: Proceedings of the 34th Annual Meeting of the Association of Computational Linguistics, Santa Cruz, CA, pp. 184–191 (1996)
19. British National Corpus, http://www.natcorp.ox.ac.uk/
20. Lodwick, K.L.: The Chinese Recorder Index: a guide to Christian Missions in Asia, 1867–1941. Scholarly Resources Inc., Wilmington (1986)
21. Noy, N.F., Fergerson, R.W., Musen, M.A.: The knowledge model of protégé-2000: Combining interoperability and flexibility. In: Dieng, R., Corby, O. (eds.) EKAW 2000. LNCS (LNAI), vol. 1937, pp. 17–32. Springer, Heidelberg (2000)
22. Yeh, J.-h., Sie, S.-h.: Common Ontology Generation with Partially Available Side Information through Similarity Propagation. In: Proceedings of the 2007 International Conference on Semantic Web and Web Services(SWWS 2007), Las Vegas, USA (June 2007)
23. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. Doklady Akademii Nauk SSSR 163(4), 845–848 (1965)