Building a large-scale testing dataset for conceptual semantic annotation of text

Xiao Wei

Shanghai Institute of Technology,
Shanghai 201418, China
and
State Key Laboratory of Management and
Control for Complex Systems,
Institute of Automation, Chinese Academy of Sciences,
Beijing 100190, China
Email: shawnwei@outlook.com

Daniel Dajun Zeng

State Key Laboratory of Management and
Control for Complex Systems,
Institute of Automation,
Chinese Academy of Sciences,
Beijing 100190, China
Email: dajun.zeng@ia.ac.cn

Xiangfeng Luo

School of Computer Engineering and Science,
Shanghai University,
Shanghai 200444, China
Email: luoxf@shu.edu.cn

Wei Wu*

Shanghai Institute of Technology,
Shanghai 201418, China
Email: weiwu@sit.edu.cn
*Corresponding author

Abstract: One major obstacle facing the research on semantic annotation is lack of large-scale testing datasets. In this paper, we develop a systematic approach to constructing such datasets. This approach is based on guided ontology auto-construction and annotation methods which use little priori domain knowledge and little user knowledge in documents. We demonstrate the efficacy of the proposed approach by developing a large-scale testing dataset using information available from MeSH and PubMed. The developed testing dataset consists of a large-scale ontology, a large-scale set of annotated documents, and the baselines to evaluate the target algorithm, which can be employed to evaluate both the ontology construction algorithms and semantic annotation algorithms.

Keywords: semantic annotation; ontology concept learning; testing dataset; evaluation baseline; ontology auto-construction; priori knowledge; evaluation parameters; guided annotation method; MeSH; PubMed.

Reference to this paper should be made as follows: Wei, X., Zeng, D.D., Luo, X. and Wu, W. (2018) ‘Building a large-scale testing dataset for conceptual semantic annotation of text’, *Int. J. Computational Science and Engineering*, Vol. 16, No. 1, pp.63–72.

Biographical notes: Xiao Wei received his BS from the Shandong University, China and PhD from the Shanghai University, China, all in Computer Science. He is currently an Associate Professor with the Shanghai Institute of Technology, Shanghai, and is a Postdoctoral Researcher with the Institute of Automation Chinese Academy of Sciences, Beijing, China. His research interests include web content analysis, semantic search, intelligent e-learning and e-commerce.

Daniel Dajun Zeng received his PhD from Carnegie Mellon University. He is a researcher in the Institution of Automation, Chinese Academy of Sciences. His main research interests are web computing, agent modelling, and security informatics.

Xiangfeng Luo received his Master's and PhD from the Hefei University of Technology, Hefei, China in 2000 and 2003, respectively. He is currently a Professor with the School of Computer Engineering and Science, Shanghai University, Shanghai, China. His main research interests include web wisdom, cognitive informatics, and text understanding.

Wei Wu received his Master's degree from South China University of Technology, Guangzhou, China. He is currently a Professor with the Shanghai Institute of Technology, Shanghai, China. His main research interests include big data and computer graphics.

1 Introduction

Conceptual semantic annotation is the basis of semantic search on text (Alani, 2003), which mainly involves two aspects of technology: the auto-construction of domain ontology based on a large set of texts and the automatic semantic annotation of text. Extensive research has been conducted concerning these two aspects (Shih et al., 2011; Zhdanova et al., 2008; Goh et al., 2011; Schutz et al., 2005; Chang et al., 2006; Liu et al., 2013; Bottoni et al., 2014), opening up application potentials of conceptual semantic annotation. Yes, major technological obstacles remain.

1 The scale of the testing datasets used by extant research is typically small. A semantic annotation algorithm needs to process large-scale text when it is meant to be used in the web environment. As such, it is critical to test such algorithms using large-scale testing datasets to validate their effectiveness and efficiency. Yet, for most researchers, no such testing datasets exist.

2 The testing datasets and evaluation standards are ad-hoc, lacking uniformity. The testing datasets used in most researches are constructed by the authors to meet the requirement of their experiments; therefore, different algorithms are evaluated by different testing datasets. This leads to serious questions of fairness and representativeness of evaluation results.

3 The construction of ontology cannot meet the real-world requirement of applications. The main methods of constructing ontologies are still largely manual or semi-manual. These methods result in highly-accurate results but with low efficiency of construction. Furthermore, such approaches cannot effectively tackle the dynamically changing data in the web environment. On the other hand, the accuracy of auto-construction of ontology needs to be improved.

This paper reports our research on constructing a large-scale testing dataset for semantic annotation of text and providing the baselines of evaluations for these datasets. The target testing dataset consists of a large-scale ontology and a large-scale set of annotated documents, which is used to evaluate both the ontology construction algorithms and semantic annotation algorithms.

To construct such a large-scale dataset, we need highly accurate ontology auto-construction algorithms and semantic annotation algorithms. In most situations, priori knowledge can improve the accuracy of algorithms (Luo et al., 2010). In this paper, we aim to use priori knowledge to guide the current ontology construction algorithms and semantic annotation algorithms to obtain higher accuracy. The method guided by priori knowledge can ensure the accuracy of the dataset. In addition, the automatic algorithms are suitable to process large amounts of data to enlarge the scale of testing dataset. The main work of the paper is as follows:

1 seek a suitable large-scale data source with priori knowledge as the basis of the target testing dataset

2 construct an accurate ontology on the dataset using ontology auto-construction methods with the guidance of priori knowledge

3 annotate the dataset by automatic semantic annotation methods with the guidance of priori knowledge

4 analyse the actual testing effects of typical algorithms as the evaluation baseline.

The rest of the paper is organised as follows. Related work is discussed in Section 2. In Section 3, we discuss how to select the data resource for the testing dataset. In Section 4, we discuss how to construct the domain ontology on the testing dataset and its evaluation parameters. Section 5 focuses on the automatic semantic annotation method on the testing dataset and its evaluation parameters. In Section 6, we introduce our testing dataset and its evaluation baselines. We conclude the paper in Section 7 by summarising the key findings and discussing future research.

2 Related work

The related work falls into two aspects, the auto-construction of domain ontology and the automatic semantic annotation.

To the first aspect, domain ontology is the basis of concept-based semantic annotation of text. Currently, highly-accurate ontologies are constructed manually, such as semantic dictionary, which has the disadvantage of slow update. Extensive researches have concerned how to construct ontology automatically to adapt to the rapid change of massive documents on the web. Shih et al. (2011) proposes a crystallising approach to enhancing domain ontology construction. Zhdanova et al. (2008) proposes a community-driven ontology construction method in social networking portals. Goh et al. (2011) proposes an automatic ontology construction method in fiction-based domain. Tarng et al. (2011) uses a virtual reality design method for learning the basic concepts of synchrotron light. These methods have improved the precision of automatic ontology construction in various degrees.

To the second aspect, based on a given domain ontology, the high precision of semantic annotation method is the foundation of all kinds of semantic services, such as document clustering, document query, and so on. Many researches have focused on how to improve the precision of semantic annotation on documents. Chen improves semantic annotation method for documents based on ontology (Chen et al., 2009). Vallet proposes an ontology-based information retrieval model (Vallet et al., 2005). Schutz presents a tool for relation extraction from text in ontology extension (Schutz, 2005). Chang proposes a query reformulation method using automatically generated query concepts from a document space (Chang et al., 2006). Kavitha uses shuffled frog leaping algorithm to improve the annotation-based document classification (Kavitha et al., 2014).

However, there are three limitations in researches of the above two aspects, which have been discussed in the introduction section in detail.

1 the scale of the testing datasets used by extant research is typically small

2 the testing datasets and evaluation standards are ad-hoc, lacking uniformity

3 the construction of ontology cannot meet the real-world requirement of applications.

Our methods in this paper are different from these researches. We develop a systematic approach on the basis of guided ontology auto-construction and annotation methods which use little priori domain knowledge and little user knowledge in documents. The proposed method is more suitable to construct a large-scale testing dataset automatically.

3 Selection of the basic data source

According to the above discussion, the basic data source should fulfil the following requirements so as to facilitate the construction of the large-scale testing dataset.

1 the scale of the date source should be large enough to build a large scale testing dataset

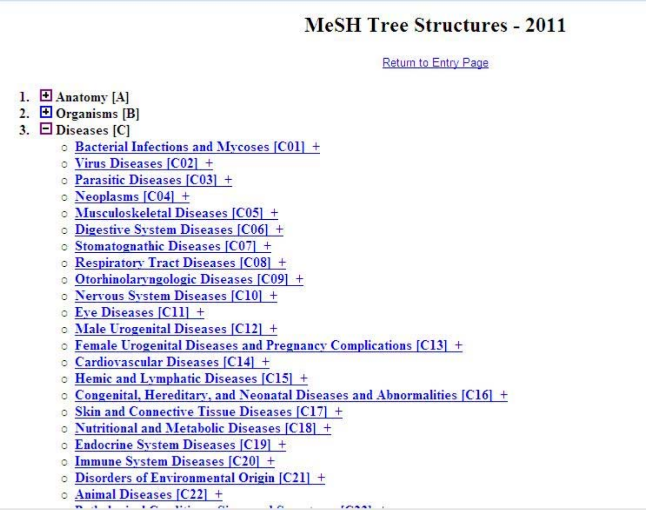
2 the documents in the data source should include priori knowledge, such as manual annotations, classification information, catalogue information, and so on, to guide the automatic semantic annotation algorithms to improve their accuracy

3 the data source should include domain knowledge, such as domain dictionary, to act as priori knowledge to improve the accuracy of ontology construction algorithm.

Based on the above mentioned, we select the medical database ‘PubMed’ and the medical thesaurus ‘MeSH’ as the basic data source of the target testing dataset.

MeSH is the most authoritative and frequently-used standard medical thesaurus, which has been used by well-known medical databases, such as PubMed, Medline, CBMdisc, as thesaurus search. MeSH is a hierarchical thesaurus, which arranges all the controlled vocabulary from the view of subject classifications and reveals the membership relations and parallel relations among them. MeSH provides a way to select words from classifications. In fact, MeSH can act as the concept tree of medical domain. A sample of the tree structure of MeSH is shown in Figure 1 (Mesh, 2014).

Figure 1 The tree structure of MeSH (see online version for colours)



MeSH Tree Structures - 2011

Return to Entry Page

1. ⊞ Anatomy [A]
2. ⊞ Organisms [B]
3. ⊟ Diseases [C]
   - ○ Bacterial Infections and Mycoses [C01] +
   - ○ Virus Diseases [C02] +
   - ○ Parasitic Diseases [C03] +
   - ○ Neoplasms [C04] +
   - ○ Musculoskeletal Diseases [C05] +
   - ○ Digestive System Diseases [C06] +
   - ○ Stomatognathic Diseases [C07] +
   - ○ Respiratory Tract Diseases [C08] +
   - ○ Otorhinolaryngologic Diseases [C09] +
   - ○ Nervous System Diseases [C10] +
   - ○ Eye Diseases [C11] +
   - ○ Male Urogenital Diseases [C12] +
   - ○ Female Urogenital Diseases and Pregnancy Complications [C13] +
   - ○ Cardiovascular Diseases [C14] +
   - ○ Hemic and Lymphatic Diseases [C15] +
   - ○ Congenital, Hereditary, and Neonatal Diseases and Abnormalities [C16] +
   - ○ Skin and Connective Tissue Diseases [C17] +
   - ○ Nutritional and Metabolic Diseases [C18] +
   - ○ Endocrine System Diseases [C19] +
   - ○ Immune System Diseases [C20] +
   - ○ Disorders of Environmental Origin [C21] +
   - ○ Animal Diseases [C22] +

PubMed (PMC) is a free full-text archive of biomedical and life science journal literature at the US National Institutes of Health’s National Library of Medicine (NLM). PubMed comprises more than 24 million citations for biomedical literature from MEDLINE, life science journals, and online books. A user, with PubMed (2014), can quickly search the entire collection of full-text articles and locate all relevant material.

A strongpoint of PubMed lies in its ability to automatically link to MeSH terms and subheadings. That is to say, each document in PubMed is annotated by Mesh terms.

The combination of MeSH and PubMed generally meets the requirements of constructing a large-scale testing dataset:

1 The combination involves abundant priori knowledge. MeSH provides thesaurus and defines the semantic

relations between terms manually. Documents in PubMed contain author's keywords, MeSH terms, etc. All of these can be thought of as priori knowledge to improve the accuracy of the testing dataset.

2 The scale of PubMed is large enough to construct a large-scale testing dataset. In addition, PubMed provides a formatted interface, which is easy to download and deal with automatically.

Despite these advantages, the combination of MeSH and PubMed cannot be used as testing dataset directly. Main problems are as follows:

1 There exist great differences between the tree structure of MeSH and the network structure of ontology, which leads to the fact that MeSH cannot be used as domain ontology directly. However, we can use terms in MeSH as seeds to mine more relevant concepts and semantic relations between concepts such as hyponymy, synonymy, similarity, etc. from documents of PubMed. Then, these concept words and semantic relations can be added to the original MeSH. After MeSH is further processed and optimised, we can obtain the domain ontology for the testing dataset.

2 There are also problems in the MeSH annotations of documents of PubMed. First, the number of annotated MeSH terms is so small that semantic information of the document cannot be fully described; Second, in the annotations, most of them are abstract concept terms of high semantic level and few of them are concept terms of low semantic level, which results in a problem that semantic information of text cannot be accurately described. If annotation terms in the documents of PubMed are viewed as priori knowledge, we can make a new annotation on the document by using concepts with different abstract levels in the newly-formed domain ontology, and then the more comprehensive semantic description and annotation on the document can be obtained.

4 Construction of domain ontology for the testing dataset and its evaluation parameters

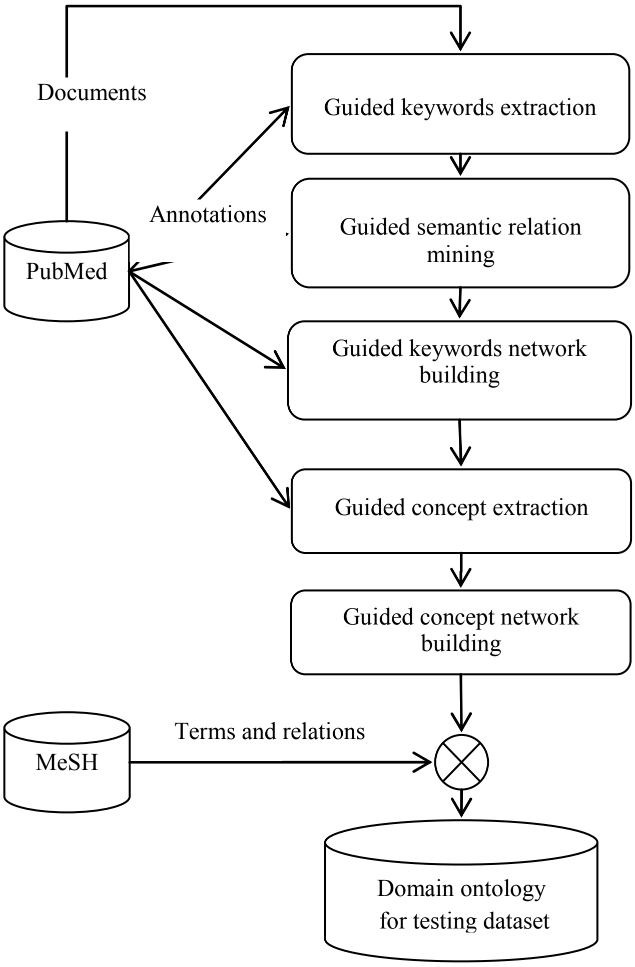
Accurate ontology is the basis of improving the accuracy of semantic annotation. Owing to the defects in the manual ontology construction such as slow update, high cost, underlying knowledge, etc., ontology construction on the large-scale dataset needs to employ auto-construction method. However, ontology auto-construction method has to face such problems as low-accuracy, low-availability. To solve the problems, this section aims to discuss how to make use of the method guided by priori knowledge to improve the accuracy of ontology auto-construction.

4.1 Ontology auto-construction method guided by priori knowledge

As discussed in Section 3, the data sources 'MeSH and PubMed' contain a certain amount of priori knowledge, which can be used to inspect and instruct auto-construction of the ontology and semantic annotation of the text so as to achieve better results than before.

In this section, we propose an auto-construction method of domain ontology with the guidance of priori knowledge, which is shown in Figure 2.

Figure 2 The process of ontology auto-construction method guided by priori knowledge


Documents
Guided keywords extraction
Annotations
Guided semantic relation mining
PubMed
Guided keywords network building
Guided concept extraction
Guided concept network building
MeSH
Terms and relations
Domain ontology for testing dataset

In the proposed method, documents provided by PubMed are regarded as a collection of text. The texts are processed by the following steps, which can be realised by some current algorithms:

1 extracting keywords (Salton et al., 1988)

2 mining the semantic relations among keywords rules (Luo et al., 2008)

3 building the semantic link network of keywords (Luo et al., 2011)

4 extracting concepts based on the keywords network (Wei et al., 2010)

5 building the semantic link network of concepts (Wei et al., 2012, 2015).

In each step, the annotations from PubMed, such as user's keywords, MeSH terms annotated in documents are used as priori knowledge to improve the current methods. As a result, we can obtain a semantic link network of concept based on the documents set. The obtained semantic link network of concept includes not only the old concepts or relations already included in MeSH but also the new concepts or relations extracted from text that have not been included in MeSH. Specially, the old and new concepts may be connected for some relations in the semantic link network of concept.

At last, using the connections between new and old concepts, the obtained semantic link network of concepts is combined and fused with terms and relations provided by Mesh to get the final domain ontology for testing dataset.

4.2 Evaluation parameters for ontology construction algorithm

The quality of ontology directly reflects the quality of the algorithm that generates it. Here, we select two parameters about the quality of ontology concept and one parameter about the stability of algorithm as the evaluation parameters for ontology construction algorithms.

4.2.1 Semantic distinguish ability of ontology concept

To concepts with different semantically abstract levels, their abilities to distinguish semantic between them are also different. Generally speaking, the higher the semantically abstract levels of the concepts, the lower the accuracy of the semantic, the lower ability to distinguish semantic between them. For example, *transportation tool* has a higher semantically abstract level than *train* or *plane.* If *transportation tool* is used to annotate *train* or *plane* appeared in some text, then these two different individuals cannot be distinguished.

Definition 1: Semantic distinguish ability of ontology concept (*SDA*) is the ability to distinguish one from another when a concept is used to annotate target objects, which can be evaluated by the accuracy of the classification of concepts.

From the perspective of text semantic annotation, SDA means when a concept is used in search, we can get the most relevant content on the semantic. For instance, in searching, when a user appoints certain semantics of a concept, it will be evaluated whether the system can provide the user with precise results of the search.

4.2.2 Semantic cover ability of ontology concept

On the contrary, the higher the semantically abstract levels of the concepts, the better the expression ability of semantics, the wider the semantic coverage. Though the precision ratio falls, the recall ratio is greatly improved.

Definition 2: Semantic cover ability of ontology concept (*SCA*) is the ability to search as much relevant information as possible in searching when the concept is used to annotate documents, which can be evaluated by the number of sub-semantic of the concept. SCA is also can be evaluated by the recall ratio of searching.

In the searching, SCA means when a user provides a concept, the system should show the user a complete semantics list of the concept for him/her to choose.

To a certain extent, there exists a contradiction between SDA and SCA. In general, from the perspective of searching, a good ontology concept should have high SCA to reach high recall of searching, and in each sub-semantic of the concept it should have high SCA to reach a high precision of searching.

To the dataset selected by the paper, the documents in PubMed have been annotated by the MeSH terms, so the precision ratio and recall ratio of each concept can be accurately calculated. Based on the precision and recall ratio, SCA and SDA of concepts can be obtained.

4.2.3 Stability of ontology construction algorithm

When ontology auto-construction algorithm is used to extract concepts, if concepts extracted at the same time are on the similar level in the hierarchy tree, namely, the semantic granularities of the concepts are similar, the method is stable. The stable method is suitable to extract the concepts with different semantic granularities by means of iteration to construct conceptual hierarchy tree of ontology. On the contrary, if concepts extracted are on different levels in the hierarchy tree, namely, the semantic granularities of the concepts are greatly different, the algorithm is not stable.

Definition 3: Stability of ontology construction algorithm (*SOC*) refers to the average of the differences of levels of the concepts extracted at a time by means of ontology construction algorithm, and can be calculated by

$$SOC = \frac{1}{n}\sum_{i=1}^{n}(Lc_i - \min(Lc)), \tag{1}$$

in which, Lc_i is a concept, c_i the level in conceptual hierarchy tree, *min*(*Lc*) the lowest level.

In the conceptual hierarchy tree, there exist differences among the semantic granularities of the concepts on the same level on the different branches. That is to say, the value of CS is affected by the conceptual hierarchy tree, but the measurement is not affected when two algorithms are evaluated and compared in the same testing dataset.

5 Semantic annotation of the testing dataset and its evaluation parameters

The documents in PubMed have been manually annotated by using the MeSH terms. Though the number of annotated terms is not large or good to show the semantic meaning of the text, these manually annotated terms are also seen as priori knowledge to instruct automatic annotation to improve the accuracy of annotation algorithm.

5.1 Semantic annotation method guided by priori knowledge

If a document d is annotated by ontology knowledge *ind,* then it is denoted by $\succ_{di}^{ind} =< ind, d, r >,$ in which r refers to relevancy of annotation between them.

Definition 4: Semantic annotation (*SA*) is the mapping from the knowledge base and text base to the annotation result, denoted by $\delta : ds \times kb \rightarrow \{\succ_{di}^{ind_m}\}.$

Vallet et al. (2005) annotates document according to the occurrence number of an instance of knowledge entity in the documents, which is a basic method of annotation. It is calculated according to the following rule.

Rule 1: The more frequently a tag word appears in a document, the higher the relevancy concerned with the document.

The relevancy can be calculated by

$$R(ind, d) = \frac{count(ind, d) \times len(ind)}{len(d)}, \tag{2}$$

in which, $R(ind,\ d)$ is the relevancy between knowledge instance *ind* and a document *d*, *count*(*ind, d*) means the frequency of knowledge instance *ind* in the document *d*, *len*(*ind*) shows the length of the tag of knowledge instance *ind*, *len*(*d*) is the length of the document *d*.

The main shortcoming of this method is to ignore the semantic environment of knowledge instance, which may result in totally wrong annotations (Chen et al., 2009).

To solve the problem, based on the annotation results of equation (2), we also consider the following rules:

Rule 2: Judge whether a concept appears in a document by the similarity between this concept and similar concepts in the document.

Using Rule 1, the frequency of a concept in a file is directly calculated, which is seen as the basis of final judgment. After a concept has successfully matched a document, semantic similarity of a concept can be further calculated to decide on whether the match is successful or not.

Denote a concept in ontology by $Co = (p_1, p_2, \ldots, p_n)$, in which n refers to the number of attribute words in concept Co, p_i the weight of attribute word in concept. The corresponding concept in the document is $Cd = (k_1, k_2, \ldots, k_m)$, in which m refers to the number of attribute words in concept Cd, k_i weight of attribute words in concept. The method of cosine similarity is employed to judge the similarity of concept *Co* and concept *Cd,* which is as followed

$$Sim(C_o, C_d) = \frac{\sum_{i=1}^{t} p_i k_i}{\sqrt{\sum_{i=1}^{n} p_i^2} \sqrt{\sum_{i=1}^{m} k_i^2}} \tag{3}$$

in which, m may not be equal to n, and t is the number of attribute words which do not repeat in *Co* and *Cd* (Salton, 1973).

Rule 2 considers context and semantics of concept. It is considered whether semantic meaning of concepts in text matches semantic meaning of concepts in ontology.

Rule 3: Keep the proportion of annotated concepts on different semantic levels.

The annotated concepts in text should come from different levels of conceptual hierarchy tree. Thus, different semantic granularities service can be offered based on the annotated materials.

Because the concepts of different semantic levels have different abilities in semantic description, the concepts used to annotate a document should have a few concepts of high level and more concepts of low level. With the decrease of semantic levels, the number of annotated concepts should gradually increase. The strong point of the method is to locate the document as quickly as possible by the concepts of high semantic level and improve recall ratio. In addition, this method does good to locate the document accurately by the concepts of low semantic level and improve precision ratio.

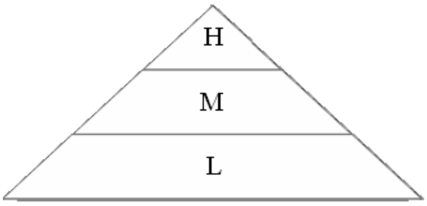
According to stability of triangle and 2/8 rule, Rule 3.1 defines the proportion of the number of concepts of different semantic levels. The proportion can ensure the semantic stability of document. It can avoid semantic uncertainty resulting from too many semantic concepts of high level, and avoid lacking globe semantic resulting from too many concepts of low semantic level.

Rule 3.1: In semantic annotation of text, the proportion of concepts from difference semantic levels should be: concepts of high semantic level to concepts of middle semantic level is two to eight, and (concepts of high semantic level + concepts of middle semantic level) to concepts of low semantic level is two to eight.

The proportion of concepts of three semantic levels is formed, shown in Figure 3, which is a triangle indicating the stability.

In order to obtain such proportions shown in Figure 3, the actual levels of all the annotated concepts in the conceptual hierarchy tree are mapped into the high, middle and low semantic levels. The specific way to realise the mapping is shown in Rule 3.2.

Figure 3 Keep the proportion of annotated concepts on different semantic levels



Notes: H: concepts of high semantic level,
M: concepts of middle semantic level,
L: concepts of low semantic level

Rule 3.2: Mapping rule of concept semantic levels. Average the *semantic levels of* conceptual hierarchy tree into three semantic levels: high, middle, low. Each annotated concept is mapped into these three semantic levels according to their actual semantic levels in conceptual hierarchy tree.

When the actual annotated concepts cannot meet the demands of the proportion in Rule 3.1, the number of annotated concepts with different semantic levels should be adapted to meet the proportion. The method of adaptation is mentioned in Rule 3.3.

Rule 3.3: Concept reservation rule. Sort the concepts by the weights of relations between the concepts with the concepts in neighbour semantic levels. The concept that has weak relevancy to other concepts should be reserved in the adaptation preferentially.

Because the concepts that have weak relevancy to other concepts are more independent in semantics, the concepts left in the collection of annotated concepts can have a stronger ability to express the semantic meaning of text.

5.2 Evaluation parameters for semantic annotation algorithm based on ontology concept

This section aims to discuss the evaluation parameters of semantic annotation algorithm based on ontology concept, perform experiments to present the effects of four annotation methods: the proposed method in this paper, MeSH annotation, and the other two methods based on Rule 1 and Rule 2 respectively.

5.2.1 Recall ratio

To a text, the recall ratio of annotation algorithm based on ontology refers to proportion of correct concepts of the annotated concepts in all the manually annotated concepts in the text, denoted as

$$recall = \frac{count(Ta \cap Tm))}{count(Tm)}, \tag{4}$$

in which *Ta* refers to concept collection obtained by annotation algorithm, *Tm* refers to concept collection obtained by manual annotation. *Ta* ∩ *Tm* refers to the intersection of concept collections *Ta* and *Tm*, which are concepts annotated correctly.

5.2.2 Precision ratio

To a text, the precision of annotation algorithm based on ontology concept refers to proportion of correct concepts in all the annotated concepts in the text, denoted as

$$precise = \frac{count(Ta \cap Tm))}{count(Ta)}, \tag{5}$$

in which, the meanings of *Ta, Tm* and others are same as (4).

5.2.3 Concept semantic level distribution

According to Rule 3, the proportion of annotated concepts belonging to different semantic levels should meet certain requirements as Figure 3 declares. Here we use concept semantic level distribution to describe the proportion of annotated concepts on different semantic levels when a text is annotated by some annotation algorithm.

To a text, the concept semantic level distribution of annotation algorithm, abbreviated as *csld*, is defined as

$$csld = \sqrt{(p_H - 0.04)^2 + (p_M - 0.16)^2 + (p_L - 0.8)^2} \tag{6}$$

in which, 0.04, 0.16 and 0.8 are the expected proportions of the concepts of high, middle, low semantic levels respectively in all annotated concepts, which are calculated according to the conditions declared in Figure 3; p_H, p_M, p_L denote the result proportions of the concepts of high, middle, low semantic levels respectively in all annotated concepts, calculated by an annotation algorithm. According to the definition of *csld*, smaller value of *csld* means the semantic level distribution of annotated concepts meets the requirement of Rule 3.1 much more.

5.3 Experiments

5.3.1 Experimental dataset

We selected three collections of documents from PubMed. Each collection includes 5,000 documents and each document contains MeSH annotation information.

The MeSH terms annotated in document and author's keywords are considered as the right concepts, which are the references used to calculate the recall and precision of the tested algorithms.

5.3.2 Annotation algorithm

Here we compare four annotation algorithms:

A0 The MeSH annotation method, directly using the MeSH annotations and author keywords in document as the annotation results, which is a kind of manually annotation method.

A1 The counting annotation method, directly based on Rule 1.

A2 The similar annotation method, directly based on Rule 2.

A3 The guided annotation method, proposed in this paper, which considers the MeSH, the annotated MeSH terms and the author keywords in the documents as priori knowledge.

5.3.3 Results

We implemented the four methods (A0, A1, A2, A3) on the three document collections (Cl, C2, C3), and calculated the precision and recall of each time. The results of precision are shown in Table 1, and the results of recall are shown in Table 2.

Table 1 Precision comparison of four annotation methods

Precision	*C1*	*C2*	*C3*	*Average*
A0	1	1	1	1
A1	0.56	0.55	0.57	0.56
A2	0.68	0.69	0.67	0.68
A3	0.85	0.86	0.85	0.85

Table 2 Recall comparison of four annotation methods

Recall	*C1*	*C2*	*C3*	*Average*
A0	1	1	1	1
A1	0.68	0.67	0.69	0.68
A2	0.78	0.77	0.77	0.77
A3	0.92	0.93	0.91	0.92

From the experimental results shown in Tables 1 and 2, it is obvious to notice the strongpoints of the semantic annotation method proposed in the paper, which are listed as follows:

1 keeping certain proportion of concept words with low semantic level does good to the specific semantic description (shown by high precision)

2 keeping concept words with high semantic level is good for summarising semantic meaning of high level (shown by high recall ratio).

6 Construction of the testing dataset

In this section, based on the MeSH and PubMed, we generate several testing datasets of different scales by means of the methods in Sections 4 and 5. Each dataset is annotated by using four methods discussed in Section 5. The values of evaluation parameters of each method are calculated which can be used as reference to evaluate other algorithms.

6.1 Steps of constructing testing datasets

1 Taking different numbers of the documents from PubMed to form four collections of text with different scales including 1,000 documents, 10,000 documents, 100,000 documents, 1,000,000 documents, and 2,000,000 documents respectively. Except the scale of 2,000,000, each scale is repeated in three different medical fields. Totally, collections of text amount to 13.

2 Getting MeSH annotation information from the documents in each collection.

3 Employing the method discussed in Section 4 to construct the ontology of each collection automatically.

4 Based on the ontology of each collection, using the four methods discussed in Section 5 to annotate the documents semantically in each collection.

5 Calculating the values of evaluation parameters as the baseline of evaluation.

6.2 The reference baselines of testing dataset

6.2.1 Reference of the precision of annotation

The precision of four methods on each testing data is calculated and shown in Table 3. The values in column 3 of Table 3 are equal to 1 because all the precision is calculated based on MeSH annotations and author keywords.

Table 3 Reference of the precision ratio of annotation

Testing dataset	*Scale*	*A0*	*A1*	*A2*	*A3*
T1	1,000	1	0.562	0.681	0.851
T2	1,000	1	0.573	0.692	0.843
T3	1,000	1	0.558	0.694	0.863
T4	10,000	1	0.532	0.681	0.865
T5	10,000	1	0.545	0.682	0.861
T6	10,000	1	0.532	0.678	0.860
T7	100,000	1	0.531	0.665	0.846
T8	100,000	1	0.522	0.673	0.842
T9	100,000	1	0.548	0.671	0.849
T10	1,000,000	1	0.506	0.651	0.842
T11	1,000,000	1	0.510	0.662	0.835
T12	1,000,000	1	0.512	0.645	0.836
T13	2,000,000	1	0.498	0.601	0.802

6.2.2 Reference of the recall of annotation

The recall ratios of four methods on each testing data are calculated and shown in Table 4.

6.2.3 Reference of concept semantic level distribution

The concept semantic level distribution of four methods on each testing data are calculated and shown in Table 5.

Table 4 Reference of the recall ratio of annotation

Testing dataset	*Scale*	*A0*	*A1*	*A2*	*A3*
T1	1,000	1	0.682	0.761	0.921
T2	1,000	1	0.694	0.762	0.912
T3	1,000	1	0.684	0.773	0.931
T4	10,000	1	0.672	0.761	0.921
T5	10,000	1	0.691	0.754	0.942
T6	10,000	1	0.686	0.762	0.916
T7	100,000	1	0.645	0.765	0.915
T8	100,000	1	0.635	0.756	0.902
T9	100,000	1	0.659	0.732	0.906
T10	1,000,000	1	0.626	0.721	0.895
T11	1,000,000	1	0.636	0.715	0.891
T12	1,000,000	1	0.659	0.726	0.881
T13	2,000,000	1	0.631	0.717	0.864

Table 5 Reference of concept semantic level distribution

Testing dataset	*Scale*	*A0*	*A1*	*A2*	*A3*
T1	1,000	0.501	0.251	0.121	0.089
T2	1,000	0.512	0.252	0.151	0.082
T3	1,000	0.511	0.263	0.153	0.081
T4	10,000	0.508	0.265	0.162	0.086
T5	10,000	0.523	0.255	0.152	0.085
T6	10,000	0.517	0.252	0.155	0.087
T7	100,000	0.521	0.265	0.173	0.081
T8	100,000	0.507	0.257	0.171	0.083
T9	100,000	0.509	0.265	0.185	0.087
T10	1,000,000	0.513	0.271	0.179	0.086
T11	1,000,000	0.518	0.275	0.177	0.089
T12	1,000,000	0.521	0.276	0.184	0.081
T13	2,000,000	0.550	0.286	0.186	0.091

7 Conclusions

To solve the problems that researches on semantic annotations lack large-scale testing datasets and the efficiency of proposed method cannot be evaluated comprehensively, this paper discusses how to build a large-scale testing dataset. The main work of this paper is summarised as follows.

1 Based on MeSH and PubMed, this paper builds a large-scale testing dataset for semantic annotation algorithms and gives the reference baseline to evaluate algorithm, which are helpful to the researches on semantic annotation.

2 This paper proposes a guided ontology auto-construction method and a guided annotation method which use little priori knowledge in the document to improve the accuracy of ontology construction and semantic annotation. The method can be directly used in the scenarios such as webpages annotation (webpages have been partly annotated), documents annotation (documents have been given some keywords).

Some potential applications derived from the proposed work are as following:

1 semantic search system of medical literatures

2 semantic dictionary of medical domain

3 semantic annotation algorithm evaluation system.

Our future focuses on building a larger-scale dataset on the selected data sources. More evaluation parameters and typical algorithms are being implemented on the dataset in order to provide more comprehensive references for evaluation.

References

Alani, H., Kim, S., Millard, D.E., Weal, M.J., Hall, W., Lewis, P.H. and Shadbolt, N.R. (2003) 'Automatic ontology-based knowledge extraction from web documents', *IEEE Intelligent Systems*, Vol. 18, No. 1, pp.14–21.

Bottoni, P., Cotroneo, A., Cuomo, M., Levialdi, S., Panizzi, E. and Passavanti, M. (2010) 'Facilitating interaction and retrieval for annotated documents', *International Journal of Computational Science and Engineering*, Vol. 5, Nos. 3–4, pp.197–206.

Chang, Y., Ounis, I. and Kim, M. (2006) 'Query reformulation using automatically generated query concepts from a document space', *Information Processing & Management*, Vol. 42, No. 2, pp.453–468.

Chen, Y.W., Li, W., Peng, X. and Zhao, W.Y. (2009) 'Improved semantic annotation method for documents based on ontology', *Journal of southeast university*, Vol. 39, No. 6, pp.1109–1113.

Goh, H.N., Kiu, C.C., Soon, L.K. and Ranaivo-Malancon, B. (2011) 'Automatic ontology construction in fiction-based domain', *International Journal of Software Engineering and Knowledge Engineering*, Vol. 21, No. 8, pp.1147–1167.

Kavitha, C., Sadasivam, G.S. and Priya, M.A. (2014) 'Annotation-based document classification using shuffled frog leaping algorithm', *International Journal of Computational Science and Engineering*, Vol. 9, No. 3, pp.215–221.

Liu, C.H., Chen, S.L. and Huang, T.Y. (2013) 'Data flow analysis and testing for OWL-S semantic web service compositions', *International Journal of Computational Science and Engineering*, Vol. 8, No. 4, pp.349–360.

Luo, X., Liang, G. and Liu, S. (2008) 'Generating associated relation between documents', in *HPCC*, pp.831–836.

Luo, X., Wei, X. and Zhang, J. (2010) 'Guided game-based learning using fuzzy cognitive maps', *IEEE Transactions on Learning Technologies*, Vol. 3, No. 4, pp.344–357.

Luo, X., Xu, Z., Yu, J. and Chen, X. (2011) 'Building association link network for semantic link on web resources', *IEEE T. Automation Science and Engineering*, Vol. 8, No. 3, pp.482–494.

MeSH (2014) [online] http://www.nlm.nih.gov/mesh/.

PubMED (2014) [online] http://www.ncbi.nlm.nih.gov/pubmed.

Salton, G. and Buckley, C. (1988) 'Term-weighting approaches in automatic text retrieval', *Information Processing & Management*, Vol. 24, No. 5, pp.513–523.

Salton, G. and Yang, C.S. (1973) 'On the specification of term values in automatic indexing', *Journal of Documentation,* Vol. 29, No. 4, pp.351–372.

Schutz, A. and Buitelaar, P. (2005) 'Relext: a tool for relation extraction from text in ontology extension', in *The Semantic Web – ISWC*, pp.593–606.

Shih, C.W., Chen, M.Y., Chu, H.C. and Chen, Y.M. (2011) 'Enhancement of domain ontology construction using a crystallizing approach', *Expert Systems with Applications*, Vol. 38, No. 6, pp.7544–7557.

Tarng, W., Lin, C.M., Liu, Y.T., Tong, Y.N. and Pan, K.Y. (2011) 'A virtual reality design for learning the basic concepts of synchrotron light', *International Journal of Computational Science and Engineering*, Vol. 6, No. 3, pp.175–184.

Vallet, D., Fermindez, M. and Castells, P. (2005) 'An ontology-based information retrieval model', in *Proceedings of the 2nd European Semantic Web Conference*, pp.455–470.

Wei, X. and Luo, X. (2010) 'Concept extraction based on association linked network', in *Proceedings of the Sixth International Conference on Semantics Knowledge and Grid (SKG)*, pp.42–49.

Wei, X., Luo, X. and Li, Q. (2012). 'Automatic facet extraction based on multidimensional semantic index', in *Proceedings of the Eighth International Conference on Semantics, Knowledge and Grids (SKG)*, pp.64–71.

Wei, X., Luo, X., Li, Q., Zhang, J. and Xu, Z. (2015). 'Online comment-based hotel quality automatic assessment using improved fuzzy comprehensive evaluation and fuzzy cognitive map', *IEEE Transactions on Fuzzy Systems*, Vol. 23, No. 1, pp.72–84.

Zhdanova, A.V. (2008) 'Community-driven ontology construction in social networking portals', *Web Intelligence and Agent Systems*, Vol. 6, No. 1, pp.93–121.