Conceptual Representing of Documents and Query Expansion Based on Ontology

Haoming Wang, Ye Guo, Xibing Shi, and Fan Yang

School of Information, Xi'an University of Finance and Economics,
Xi'an, Shaanxi 710100, P.R. China
hmwang@mail.xaufe.edu.cn, shiny_yang0000@sina.com,
{guoyexinxi,xbshine}@126.com

Abstract. In vector space model, a document is represented by words. As the new words appear dramatically in the Internet era, this kind of method draws back the IR systems performance. This paper puts forward a new approach to present the concepts, query expressions, and documents based on the ontology. The approach has two levels, the Word-Concept level and the Concept-Document level. In the first level, the transition probability matrix is constructed by using the appearing times of word-word pairs in documents. The biggest eigenvector of matrix is computed, and it reflects the importance of words to the concept. In the second level, the distance matrix is constructed by using the distance between words in a given ontology, and the average variance value of elements is computed. It reflects the relevance of documents to concepts. In the last section, the query expansion is discussed by using the personal information profile of the user. It is proofed to be more effective than previous one.

Keywords: Ontology, Word-Concept level, Concept-Document level, Relevance Computing, Personal Information Profile.

1 Introduction

With the explosive growth of information in Internet, search engine (SE) is used in information retrieval (IR). Users of IR Systems expect to find the most relevant items to a certain query. Generally, the IR system does not feed back an ideal behavior. They feed back much more results to users, and users have to spend a considerable time to find these items which are really relevant to their initial queries.

One of the reasons is that the system does not know what the user wants to get actually. The search engine separates the terms in the query expression, and computing the page value by using the contents or the page-links or the both of the pages. The $top - n$ pages are feed back to the users. This kind of searching method neglects the relevant documents that do not contain the index terms which are specified in the users queries. In order to improve the effect of retrieval, the specific domain knowledge should be added to the queries.

F.L. Wang et al. (Eds.): WISM 2012, LNCS 7529, pp. 492–499, 2012.

Ontology is a conceptualization of domain knowledge. It is a concept set with the human understandable, machine readable format. It consists entities, attributes, relationship, and axioms. For an ontology based information retrieval system, when the user input the query expression, the system tries to insert the ontology knowledge to enhance to query expression in order to increase the probability of relevancy. In the concept level, documents having very different vocabularies could be similar in subject and, similarly, documents having similar vocabularies may be topically very different.

This paper is organized as follows: Section 2 introduces the concepts of Ontology and Query Expansion. Section 3 discusses new approach combining words, concepts and documents in two levels. Section 4 presents the methods of query expansion. Finally, a summary of this paper and directions for future work are discussed in Section 5.

2 Related Works

2.1 Ontology

In the traditional IR approaches, documents and query expression are represented as a vector of terms simply. One of the examples is VSM. The relevance of the document and the query is computed by the cosine distance between two vectors. This approach does not require any extraction or annotation phases. Therefore, it is easy to implement, however, the precision value is relatively low.

Ontology is explicit representations of a shared conceptualization, i.e., an abstract, simplified view of a shared domain of discourse. More formally, an ontology defines the vocabulary of a problem domain, and a set of constraints (axioms and rules) on how terms can be combined to model specific domains. An ontology is typically structured as a set of definitions of concepts and relations between these concepts. Ontology is machine-processable, and they also provide the semantic context by adding semantic information to models, thereby enabling natural language processing, reasoning capabilities, domain enrichment, domain validation, etc [1,2].

2.2 WordNet

WordNet, a manually constructed electronic lexical database for English, was conceived in 1986 at Princeton University, where it continues to be developed. WordNet is a large semantic network interlinking words and groups of words by means of lexical and conceptual relations represented by labeled arcs. WordNets building blocks are synonym sets (synsets), unordered sets of cognitively synonymous words and phrases. Each member of a given synset expresses the same concept, though not all synset members are interchangeable in all contexts. Each synset has an unique identifier (SynsetID). Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser [3].

In this paper, we select one ontology from the WordNet to discuss our approach.

2.3 Conceptual Representation

In VSM, the term is the word of the document normally. Due to the ambiguity and the limited expressiveness of single word, it is difficult to decide which word is more important for the document.

One way of improving the quality of similarity search is Term FrequencyInverse Document Frequency (TFIDF). The main idea of TFIDF is the more vocabulary entry in document set, the lower separate ability of document property, and then the weight value is small. On the other hand, the higher frequency for a certain vocabulary entry in a document, the higher separate ability, and then the weight value is big. The method is widely used in selecting text feature. But it has many disadvantages too. First, the method undervalues that this term can represent the characteristic of the documents of this class if it only frequently appears in the documents belongs to the same class while infrequently in the documents of the other class. Second TFIDF neglects the relations between the feature and the class [4].

The another way is Latent Semantic Indexing (LSI). The most improvement is mapping the document from the original set of words to a concept space. Unfortunately, LSI maps the data into a domain in which it is not possible to provide effective indexing techniques. Instead, conceptual indexing permits to describe documents by using concepts that are unique and abstract human understandable notions. After that, several approaches, based on different techniques, have been proposed for conceptual indexing.

One of the well-known mechanism for conceptual representation is conceptual graph (CG). In Ref. [5], two ontologies are implemented based on CGs: the Tendered Structure and the abstract domain ontology. And, the authors first survey the indexing and retrieving techniques in CG literatures by using these ontologies.

2.4 Query Expansion

In an IR system, the user inputs his query expression for his requirements. Normally, the query expression is not clear enough to let the IR system understands what the user wants. Query expansion is one of the ways to solve this problem [6].

Query expansion technology was brought forward in Ref. [7]. It expands a query expression with the addition of terms that are semantically correlated with the original terms of the query. Several works demonstrated the performance of IR system was improved by using it. As the terms, which are added to the query, play a decision rule in the query process, they should be selected carefully. Experimental results show that the incorrect choice of terms might harm the retrieval process by drifting it away from the optimal correct answer [8].

3 Document Representing

There are three tasks for the documents representing based on the concept:

(1) Labeling the terms in the document. The document consists of terms, and most of them should be belonged to one or more concepts. Some of the terms are not so closed to the concepts, and they are omitted in this step. The documents are represented by the remain terms.
(2) Computing the relevance. There are many terms in a concept. There is no experimental result shows that some terms are always more important than others. The importance of the term is different in different concept. For the document, we want to get a term list by the importance decrease order to the concepts.
(3) Deciding the attribution of the document to the concept. For a given document, it may have relation with two or more concepts, which concept the document belongs to finally? Sometime, we need to answer the question that if two documents had same terms but different term orders, do they have same importance for a query or in a concept?

In the following, we construct a new approach with two-level structures. The first level, called Word-Concept (W-C) level, reflects the relation between the words and the concepts. And the second level, called Concept-Document (C-D) level, reflects the relation between concepts and documents.

3.1 Labeling Document

In our discussion, the first task is labeling the terms in a document to the concept. The method of labeling is referencing the WordNet. For the ontology and the document, we can assume the facts:

(1) An ontology is a very large set of concepts, and there are several hundreds of concepts in it. Each concept is consisted of many terms and the relations between the terms, meanwhile each term may belong to more than one concept.
(2) For a document, it is impossible to include all of the terms in a special concept or a ontology. In other words, it is impossible that all of the terms in a document are belonged to one concept.
(3) Assuming the term d is one of the terms of a document $D(d \in D)$, d can be labeled to concept C_1 or C_2 according to the term-list of the concepts. Which concept should be selected for the term d finally?

The Word-Concept level of the new approach can be described as :

(1) Constructing the matrix UC for each concept C, it is:

$$\mathbf{UC} = \begin{pmatrix} uc_{11} & uc_{12} & \dots & uc_{1n} \\ uc_{21} & uc_{22} & \dots & uc_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ uc_{n1} & uc_{n2} & \dots & uc_{nn} \end{pmatrix}$$

Where the element $uc_{ij}(i \neq j)$ is the times which word d_i and d_j appear synchronously in a paragraph, and uc_{ii} is the times which word d_i appears in a paragraph by itself. In the beginning, all of the elements are 0.

(2) Scanning the document D. We label the words to the different concepts. If a word was belong to two or more concepts, it was labeled to each of the concepts. After the scanning, we count the times, of which word d_i and d_j appear synchronously, and replace the uc_{ij} with the value.

(3) Dealing with the matrix UC. If the column i is all 0, it means the word d_i does not appear in document D. The column i and row i of this matrix should be deleted.

It is obviously that the UC is symmetric matrix. In order to decrease the amount of computation, we set a threshold for value of elements. Deleted the rows and columns synchronously, the matrix keeps the characters of symmetric.

The document D may have relevance with concepts $C_1, C_2, \ldots, C_k$. We denote the relevance by matrix $UC_p, p \in [1, k]$. In the following distribution , we indicate the matrix $UC_p, p \in [1, k]$ with Q for convenience.

3.2 Computing Relevance

The elements $q_{ij}(i, j \in [1, n])$ of matrix Q responds to the times of word pair d_i-d_j appeared in the same paragraph in a document D. Normalizing the matrix Q, we explain it as:

We have a word set $D = \{d_1, d_2, \cdots, d_n\}$, and we name each word in the set with the state. The process starts in one of these states and moves successively to another. Each moving is called a step. If the chain is currently in state d_i, then it moves to state d_j at the next step with a probability denoted by q_{ij} , and the probability does not depended upon which states the chain was in before . The word set $D = \{d_1, d_2, \cdots, d_n\}$ can be regarded as Markov Chain. The Q is row-stochastic matrix, and the elements q_{ij} is transition probabilities.

According to the Chapman-Kolmogorov equation [9] and characters of Markov chain [10], the $n - step$ transition matrix can be obtained by multiplying the matrix Q by itself n times.

In the Q , the elements are connected to others, and the matrix cannot be divided into two parts. So the Q is irreducible. Meanwhile the Q is aperiodic too. The Perron-Frobenius theorem guarantees the equation $x^{(k+1)} = Q^T x^{(k)}$ (for the eigensystem $Q^T x = x$)converges to the principal eigenvector with eigenvalue 1, and there is a real, positive, and the biggest eigenvector [11].

The biggest eigenvector means the importance of word d_i to the concept C.

3.3 Deciding Affiliation

In the above, we compute the importance of a word to a concept. In this section, we will discuss the relevance of the document to the concept. The relevance will be computed in Concept-Document(C-D) level of our new approach.

Assuming we have two documents D_1 and D_2 and a concept C, how can we consider the D_1 has more or less relevance than the D_2 to the C?

The graph G discussed previously is consisted by the nodes and the links. The node represents the word and the link represents the relation. According to the definition of ontology, there are four kinds of relation between the words, such as $part-of$, $kind-of$,$instancd-of$ and $attribute-of$. We define the relation between the words as,

Define 1. *Assuming w_i and w_j are the nodes of graph G. If w_i does not connect to w_j directly, there is a path from w_i to w_j. The distance between them is the minimum of the steps from w_i to w_j.*

$$distance(w_i, w_j) = Min(n|w_i \to w_1 \to w_2 \to \cdots \to w_n \to w_j)$$

Define 2. *If w_i connected to w_j directly, the distance between them is,*

$$distance(w_i, w_j) = \begin{cases} 1\ Rela(w_i,w_j) \in \{part-of, attribute-of\} \\ 2\ Rela(w_i,w_j) \in \{instance-of, kind-of\} \end{cases}$$

Here $Rele(w_i, w_j)$ is the one of the four relations between the words in a ontology.

According to the $Define1$ and $Define2$, we construct the distance matrix $Dis(C, D)$, which represents the distance of the document D to the concept C.

$$\mathbf{Dis(C, D)} = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ 0 & d_{22} & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & d_{nn} \end{pmatrix}$$

In $Dis(C, D)$, column's order can be exchange in order to keep the column i is the hypernym of column j when $i < j$.

The sum of the row i, named it with $Dis(i) = \sum_j d_{ij}, j \in [1, n]$, means the ability of word w_i representing the concept. In general, the more the $Dis(i)$, the much irrelevance of the word w_i to the concept.

So, assuming the matrix $Dis(C, D_1)$ and $Dis(C, D_2)$ represent the relevance of document D_1 and D_2 to the concept C respectively, we compute the distance respectively just as fellows,

(1) Computing the Average Variance of each rows $AV_i, i \in [1, n]$;
(2) Sum the Average Variance value $V = \sum AV_i, i \in [1, n]$.

Hence, we get two Average Variance values V_1 and V_2 for the document D_1 and D_2 to the concept C respectively. We consider that the document with the less Average Variance of V_1 and V_2 has much relevance with the C.

4 Query Expansion

Search Engine (SE) plays the important role in finding information in Internet. The user inputs the query sentence to SE, and the most important thing for SE

is to know what the user want to get exactly. In normal, the query sentence is not detail enough to be used to feed back the satisfactory results to the user. Query expansion is used to solve this problem.

There are many ways to expand the query sentence. But it is difficult to expand it without any other help, such as the domain information, surfing history or log records. In our approach, the user is requested to register for the personalized service. The personal information is used to construct the Personal Information Profile (PIP). After the IR system feed back the results to the user, he checks the results and estimates them. The IR system refine the PIP according to the estimation. In the constructing of PIP, the ontology play a key role. By using ontology, we can enrich the implication of query and to enhance the search capabilities of existing web searching systems. The method of expanding can be described as,

(1) Splitting the query to words and marking them in the ontology words pool. The weight of the word plus 1 for each time appeared in the query. It is obviously that the more times the word appear in the query, the more weight it is in the ontology words pool.
(2) Selecting the concept, which the query words are involved in, we order the words belong to the concept just as following steps,
   (a) Ordering two word-lists. The first one is that the words order by the relevance, which are computed in W-C level. We named it as,

$$M(w_{i1}, w_{i2}, ..., w_{im}).$$

   The second one is that the words order by the appearance times in user's query in a given period. We named it as,

$$N(w_{j1}, w_{j2}, ..., w_{jn}).$$

   (b) Setting the final word list as,

$$P = \alpha M(w_{i1}, w_{i2}, ..., w_{im}) + (1 - \alpha)N(w_{j1}, w_{j2}, ..., w_{jn}), \alpha \in (0, 1).$$

   (c) Setting the threshold, and selecting the $top-R$ words. The $top-R$ words have much relevance with the words appeared in the query sentence.

   In general, it can be imaged that the effect of this way is not ideal in the beginning as the limited of the words in query sentence. With the times of query input increased, the accuracy will be better.
(3) The R words selected in the last step will be submitted to the SE, and SE feed backs the results to user according to the these words. The user reviews the results, and he presents his owner opinion for the retrieval results. The opinion will be used to refine the parameter $\alpha \in (0, 1)$ in the formula.

5 Conclusion

The paper introduces the concepts of ontology, query expansion, and representing the document by using the ontology. We construct a new approach with two

levels, the Word-Concept level and Concept-Document level, which reflects the relation among the words, concepts, documents and queries. By computing the biggest eigenvector of words matrix to determine the relevance of words appeared in document to concepts, and computing the average variance to determine the distance of document to the concept. By constructing the Personal Information Profile(PIP) of user to expand the query sentence. According to the forecast, the feedback results will be fine than before.

Acknowledgment. This work was supported by Scientific Research Program Funded by Shaanxi Provincial Education Department, P.R.China (Program No.09JK440), and Natural Science Foundation of Shaanxi Province of China (Program No.2012JM8034).

References

1. Kara, S., Alan, O., Sabuncu, O., Akpınar, S., Cicekli, N.K., Alpaslan, F.N.: An ontology-based retrieval system using semantic indexing. Information Systems 37(4), 294–305 (2012)
2. Kang, X., Li, D., Wang, S.: Research on domain ontology in different granulations based on concept lattice. Knowledge-Based Systems 27, 152–161 (2012)
3. Dragoni, M., da Costa Pereira, C., Tettamanzi, A.G.: A conceptual representation of documents and queries for information retrieval systems by using light ontologies. Expert Systems with Applications (2012) 10.1016/j.eswa.2012.01.188
4. Qu, S., Wang, S., Zou, Y.: Improvement of text feature selection method based on tfidf. In: International Seminar on Future Information Technology and Management Engineering, FITME 2008, pp. 79–81 (November 2008)
5. Kayed, A., Colomb, R.M.: Using ontologies to index conceptual structures for tendering automation. In: Proceedings of the 13th Australasian Database Conference, ADC 2002, vol. 5, pp. 95–101. Australian Computer Society, Inc., Darlinghurst (2002)
6. Kim, M.-C., Choi, K.-S.: A comparison of collocation-based similarity measures in query expansion. Information Processing and Management 35(1), 19–30 (1999)
7. Efthimiadis, E.N.: Query expansion. Annual Review of Information Science and Technology 31, 121–187 (1996)
8. Cronen-townsend, S., Zhou, Y., Croft, W.B.: A framework for selective query expansion. In: Proceedings of Thirteenth International Conference on Information and Knowledge Management, pp. 236–237. Press (2004)
9. Gardiner, C.: Stochastic Methods: A Handbook for the Natural and Social Sciences. Springer Series in Synergetics. Springer (2009)
10. Mian, R., Khan, S.: Markov Chain. VDM Verlag Dr. Muller (2010)
11. Serre, D.: Matrices: theory and applications. Graduate texts in mathematics. Springer (2010)