WCD-New Approach Combining Words, Concepts and Documents Based on Ontology

Haoming Wang, Ye Guo, and Xibing Shi

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

Abstract. In traditional Information Retrieval (IR) system, the document is represented by the set of words or terms. If the words or terms are regarded as the components of a vector, the model is called the vector space model (VSM). VSM has been widely used in IR systems in recently decades. As the the new words appear dramatically in the Internet era, the amount of computation is very large and it draws back the IR system's performance. This paper puts forward a new approach according to the relations among the words, concepts and the document by using the concept of the ontology. The new approach has two levels, the Word-Concept (WC) level and the Concept-Document (CD) level. In the WC level, the transition probability matrix is constructed by using the word-word pairs appeared in the same paragraph, and the biggest eigenvector of matrix is computed. The eigenvector reflects the importance of the word to the concept. In the CD level, the distance matrix is constructed by using the distance between words in the concept, and the average variance values of elements is computed. The value determines the relevance of the document to the concept. In order to expand the query sentence, the Personal Information Profile (PIP) of the user is defined by using the query history of the user. It is proofed to be more effective than previous one.

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

1 Introduction

In Internet era, search engine (SE) is used to help the user to find the information in Internet. The user always expects to find the most relevant information to his query. The recall and precision are used to test the effects of retrieval results. Normally, the SE cannot always feed back an ideal result list as the various reasons. The user has to spend a considerable time to find the useful information in the feedback list. The main reason of causing the poor results is that the SE does not really know what the user wants to get. The SE splits the query sentence

Z. Li et al. (Eds.): ISICA 2012, CCIS 316, pp. 449–458, 2012.
© Springer-Verlag Berlin Heidelberg 2012

into terms, and computing the page value by using the contents or the page-links of the pages or the both.

The first kind of SE works on the contents of pages or documents. The typical example is vector space model (VSM). This model regards the document as a vector which is consisted of many component, and each component is a term. In the most time, the term is the word. So, the more the different terms in a document, the more the dimensions of the vector. This method works well for traditional documents, but the performance drops significant when applied to the web pages. The main reason is that the web pages are not organized as strict as the documents do. Web pages contain much irrelevant information, and the useful information is submerged in. On the other hand, the VSM model neglects the relevant pages that do not contain the index terms which are specified in the user's queries.

The second one takes the hyperlink structures of web pages into account in order to improve the performance. The examples are Pagerank and HITS. They are applied to Google and the CLEVER project respectively. The Google search engine is based on the popular Pagerank algorithm first introduced by Brin and Page in[1]. Considering the pages and the links as a graph $G = P(Page, Link)$, we can describe the graph by using the adjacency matrix. Computing the eigenvector which represents the value of the pages of the matrix, the Google feedbacks the top $-n$ pages to the user. The [2,3] introduced the algorithm, but they did not introduce the deep technologies of the SE, which they are the trade secrets. The [4] shows the result that Google cannot get the high level of precision and recall if they compute the page value just by using the links between the pages. They must have other technologies to improve the results of retrieval.

Normally, in order to improve the effect of retrieval, the specific domain knowledge should be added to the queries, which is called query expansion. Ontology is one of the knowledge which can be used to expand the connotation of the query.

Ontology is a conceptualization of a domain knowledge. It is a concept set with the machine readable, human understandable format. It consists entities, attributes, relationship, and axioms. Those elements construct the network in order to present the knowledge of a special domain. The network is checked by the experts of the domain in order to guarantee the concepts represent the knowledge of the domain indeed.

For an ontology-based information retrieval system, it tries to insert the ontology knowledge to the query expression in order to enhance the ability of representing. In the conceptual level, documents having very different vocabularies could be similar in subject and, similarly, documents having similar vocabularies may be topically very different.

The paper is organized as follows: Section 2 introduces the related concept of ontology and query expansion. Section 3 discusses the relevance computing of words to concepts and documents to concepts. Section 4 puts forward the method of query expansion. In section 5, a summary of the paper and directions for future work are discussed.

2 Related Work

In this section, the basic concepts, such as ontology, conceptual representing, document representing, query expansion, and WordNet are introduced.

2.1 Conceptual Representation

If a document is considered as a set of words, the relation between words is neglected. Due to the ambiguity and the limitation of the ability of expressing of the single word, it is difficult to decide which word is more important than others for the document. It is obviously that the importance of the word is decided by the document or set of documents which the word appeared. There is no experimental result shows that a word is always important that the others. So, we try to discuss the importance of words in an ontology or a concept environment.

One way of deciding which word is more important than others is Term Frequency-Inverse 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. This method is widely used in selecting text feature[5]. But it has many disadvantages. 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 or the terms[6].

The another way is Latent Semantic Indexing (LSI). The most improvement is mapping the documents 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 [7], 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.

2.2 Ontology

According to [8,9], an ontology is considered as a set of definitions of concepts and relations between these concepts with typically structure. It provides the semantic context by adding semantic information to models. It is machine-processable, and it can be used in natural language processing, reasoning capabilities, domain enrichment, domain validation, etc.

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.

In the traditional IR system based on VSM, documents and queries are simply represented as a vector with the term weight, and the similarity is computed by the cosine distance between the vectors. This approach does not require any extraction or annotation phases. Therefore, it is easy to implement, however, the precision values are relatively low. Compared with the traditional approach, the new one expands the query by using the knowledge of relative domains. It is hoped to improve the precision and the recall.

2.3 Query Expansion

Normally, the query expression is consisted of 3 to 7 words. The information including in those words is not clear enough to let the IR system know what the user wants really. The IR system has to feed back the much more results to the user. This causes the lower precision and recall. Query expansion is one of the ways to solve this problem[10].

Query expansion technology was brought forward in [11]. It consists of expanding a query 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[12].

There are many ways to select the words to add to the query sentence. One way is to add the synonyms of that words appeared in the query sentence. The task of defining the relation between the words is urgent. Further more, it is better to consider that in semantic level. In this paper, we select the synonyms in conceptual level of an ontology.

3 Document Representing

The document is an objective reality, and it can be dealt with by many kinds of ways[13]. The VSM discussed above is one of the methods. VSM regards the document as a vector combined with the words. Each word is regarded as a component of the vector, and the relations between the words are neglected.

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

We discuss the W-C level first. For a document, there are three tasks in representing it based on the concept:

1. Marking the words in the document. The document consists of words, most of them should be belonged to one or more concepts. Some of the words are not so closed to the concept, and they are omitted in this step. The document is represented by the remain words.

2. Computing the relevance of word-document and document-concept. For a given concept, no experimental result shows that some words are always more important than others. The importance of the word is different in different documents. We want to get a word 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 relevance with two or more concepts. In the other side, there are many documents have relevance to a given concept. Sometime, we need to answer the question that if two documents had same words set but different words order list, which one is more important for a query or a concept?

3.1 Marking Words

In our discussion, the first task is marking the words in a document. For the concept 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 words, meanwhile each word may belong to more than one concept.
2. For a document, which is a instance for a given topic, it is impossible to include all of the words in a special concept or a ontology. In other words, it is impossible that all of the words in a document are belonged to one concept or one ontology.
3. Assuming the word d is one of the words of a document $D(d \in D)$, d can be labeled to concept C_1 or C_2 or both according to the term-list of the concepts.

We select the concepts in WordNet as the working level. The Word-Concept level of the new approach can be described as:

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

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

   Where the element uc_{ij} is the times which word d_i and d_j appear synchronously in a paragraph, and $u_s c_{ii}$ is the times which word d_i appears in a paragraph by itself. In the beginning, all of the elements is 0.
2. Scanning the document D from the first word to the end, we mark the words to the different concepts. If a word is belong to two or more concepts, marking it to each of the concepts. After the scanning, we sum the times, of which word d_i and d_j appear synchronously in the same paragraph. The value of element uc_{ij} in matrix UC is updated by it.

3. Dealing with the matrix UC. If the column i is all zero, it means the word d_i never appear in document D. The column i and row i of this matrix should be deleted.

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

The document D may have relevance with concepts $C_1, C_2, ..., C_k$, $k \in [1, n]$. We denote the relevance by matrix UC_p, $p \in [1, n]$. In the following section, we indicate the matrix UC_p, $p \in [1, n]$ 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. The $row(i)$ means the probability of word d_i and the word d_j, $j \in [1, n]$ appear at the same time in this document. Normalizing the matrix Q, we explain it as:

We have a set of words $D = \{d_1, d_2, ..., d_n\}$, and we name each word with the state. The process starts in one of these states and moves successively from one state to another. Each move 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 this probability does not depend upon which states the chain was in before. The word set $D = \{d_1, d_2, ..., d_n\}$ can be regarded as Markov Chain. The matrix Q is row-stochastic matrix, and the elements q_{ij} is transition probabilities.

According to the Chapman-Kolmogorov equation[14],

$$q_{i_1,...,i_{n-1}}(f_1, ..., f_{n-1}) = \int_{-\infty}^{+infty} q_{i_1,...,i_n}(f_1, ..., f_n) df_n \tag{1}$$

For the Markov chains, we can get[15],

$$q_{ij}^{n+m} = \sum_{k=0}^{\infty} q_{ik}^n q_{kj}^m (n, m \geq 0, \forall i, \forall j) \tag{2}$$

If we let $Q^{(n)}$ denote the matrix of $n - step$ transition probabilities q_{ij}^n, then we can asserts that:

$$\begin{aligned} &Q^{(n+m)} = Q^{(n)} \cdot Q^{(m)} \\ &Q^{(2)} = Q^{(1)} \cdot Q^{(1)} = Q \cdot Q = Q^2 \\ &Q^{(n)} = Q^{(n-1+1)} = Q^{(n-1)} \cdot Q^{(1)} = Q^{n-1} \cdot Q = Q^n \end{aligned} \tag{3}$$

That is, the $n-step$ transition matrix can be obtained by multiplying the matrix Q by itself n times.

The elements of the matrix Q 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[16].

Because Q corresponds to the stochastic transition matrix over the graph G, the stationary probability distribution over all words induced by a random selection of words on document D can be defined as a limiting solution of the iterative process:

$$x_j^{(k+1)} = \sum_i Q'_{ij} x_i^{(k)} = \sum_{i \to j} x_i^{(k)} / deg(i) \tag{4}$$

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

3.3 Deciding Affiliation

In this section, we will discuss the C-D level, which is the relevance of the concept to the document.

Assuming the relevance of two sets, which come from the documents D_1 and D_2 respectively, to the concept C have been computed, we should decide which document has much importance to the concept?

According to the definition of ontology, there are four kinds of relation between the words, $part-of$, $kind-of$,$instancd-of$, and $attribute-of$. We define the distance between the words as,

Define 1. *Assuming w_i 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.*

$$distance(w_i, w_j) = Min(n | w_i \to w_1 \to w_2 \to ... \to w_n \to w_j) \tag{5}$$

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

$$distance(w_i, w_j) = \begin{cases} 1 \ (if \ \ Rela(w_i, w_j) \in \{part-of, attribute-of\}) \\ 2 \ (if \ \ Rela(w_i, w_j) \in \{instance-of, kind-of\}) \end{cases} \tag{6}$$

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

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

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

In $Dis(C, D)$, $d_{ij}(i, j \in [1, n])$ is the distance between word w_i and w_j. The word $w_k(k \in [1, n])$ is one of the words in set D, which is discussed in section 3.2. In order to reduce the amount of computation, we sort the words in decrease order by the importance and set the thresholds. We deal with the Top-n words only.

The column's order of $Dis(C, D)$ 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 $Dis(i) = \sum_j dis_{ij}, j \in [1, n]$, means the degree of words representing the concept or ontology. In general, the more the sum, the much irrelevance of the words to the concept or ontology.

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 follows:

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.

Section 3.1, Section 3.2 and Section 3.3 introduce the our new approach, combining the word, concept, and document based on ontology.

4 Query Expansion

Search Engine (SE) plays the important role in finding information in Internet. The user inputs the query sentence to SE, and he hope the SE can feed back the results what he want to get exactly. In normal, the query sentence is not clear enough to let the SE know what the question is. Query expansion is used to solve this problem.

As we know, it is difficult to expand the user's query sentence without any other help, such as the domain information, surfing history or log records. In this paper, we require the user to register if he want to get the personalized service. The personal information of the user is used to construct the Personal Information Profile (PIP). After the IR system feeds back the results, he checks the results and estimates them. The IR system re

ne the PIP according to the estimation. The PIP works as a filter between the user and the feedback results.

By using the PIP, the method for query expansion can be described as,

1. Splitting the query to words and marking them in the domain 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 domain words pool.
2. Selecting the concept and the words involved in the concept according to the user's PIP, 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,

$$M(wi) = M(w_{i1}, w_{i2}, ..., w_{im}) \tag{7}$$

The second one is that the words order by the appearance in the domain words pool in a given period. We named it as,

$$N(wj) = N(w_{j1}, w_{j2}, ..., w_{jn}) \tag{8}$$

(b) Setting the final word list according to the Eq.(7) and Eq.(8) as,

$$P(M, N) = \alpha M(wi) + (1 - \alpha)N(wj), \alpha \in (0, 1) \tag{9}$$

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

3. The R words selected in the last step will be submitted to the SE, and SE feedbacks 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.

In the step 2 of query expansion, the amount of computation is very huge because we have to consider the words in whole concept level. With the time going, if we could focus on the some of the words in the concept, and those words have more relevance than the others, we can reduce the amount of computation. So, 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.

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 relevance among the words, concepts, documents and queries. By computing the biggest eigenvector of words matrix to determine the relevance of words to the concepts, and computing the average variance to determine the distance of document to the concept. In the last paragraph, we introduce the way to expand the query sentence by constructing the Personal Information Profile (PIP) of user. According to the forecast, the feedback results will be fine than before.

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

References

1. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. In: Proceedings of the 7th International Conference on World Wide Web, pp. 107–117 (1998)

2. Bianchini, M., Gori, M., Scarselli, F.: Inside pagerank. ACM Transactions on Internet Technology 5(1), 92–128 (2005)
3. Altman, A., Tennenholtz, M.: Ranking systems: the pagerank axioms. In: Proceedings of the 6th ACM Conference on Electronic Commerce, EC 2005, pp. 1–8. ACM, New York (2005)
4. Wang, H.-m., Rajman, M., Guo, Y., Feng, B.-q.: NewPR-Combining TFIDF with Pagerank. In: Kollias, S.D., Stafylopatis, A., Duch, W., Oja, E. (eds.) ICANN 2006. LNCS, vol. 4132, pp. 932–942. Springer, Heidelberg (2006)
5. Jones, K.S.: A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 28, 11–21 (1972)
6. 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 (2008)
7. 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)
8. Kara, S., Alan, O., Sabuncu, O., Akpnar, S., Cicekli, N.K., Alpaslan, F.N.: An ontology-based retrieval system using semantic indexing. Information Systems 37(4), 294–305 (2012)
9. Kang, X., Li, D., Wang, S.: Research on domain ontology in different granulations based on concept lattice. Knowledge-Based Systems 27, 152–161 (2012)
10. Myoung-Cheol Kima, K.S.C.: A comparison of collocation-based similarity measures in query expansion. Information Processing and Management 35(1), 19–30 (1999)
11. Efthimiadis, E.N.: Query expansion. Annual Review of Information Science and Technology 31, 121–187 (1996)
12. 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)
13. Wu, C.C., Chou, C.H., Chang, F.: A machine-learning approach for analyzing document layout structures with two reading orders. Pattern Recognition 41(10), 3200–3213 (2008)
14. Gardiner, C.: Stochastic Methods: A Handbook for the Natural and Social Sciences. Springer Series in Synergetics. Springer (2009)
15. Mian, R., Khan, S.: Markov Chain. VDM Verlag Dr Muller (2010)
16. Serre, D.: Matrices: theory and applications. Graduate texts in mathematics. Springer (2010)