Research of the Conceptual Representing of Documents based on Light Ontology

Haoming WANG*, Ye GUO†, Jianting LI‡, and Xibing SHI§

School of information, Xi'an University of Finance and Economics
Xi'an, 710100, Shaanxi , P.R.China
*hmwang@mail.xaufe.edu.cn
†guoyexinxi@126.com
‡jianting.li@gmail.com
§xbshine@126.com

***Abstract*—The traditional method of presenting a document in an Information Retrieval(IR) system is based on terms. As the the new words appear dramatically in the Internet era, this kind of method draws back the IR system's performance. This paper puts forward an approach by using the concepts of the ontology to present the documents. Constructing the Word-Concept(W-C) model and Concept-Document(C-D) model, we compute the relevance of word-word , word-concept and concept-document. They are used to determine which page is most relevant to the query. It is proofed to be more effective than prevenient.**

I. Introduction

In the traditional Vector Space Model(VSM) approach, a document is presented by a serial of terms. In the most time, a term means a word and all of the different words in the document consist a vector, with which to represent the document. This method gained the success in document classification and relax domain [1], [2] .This model assumes two condition, the first one is that the words appear in a document is limit, that means the cost of computing can be controlled. The second one is that the words in a document is independent and there is no relation among them. But now, in the Internet era, the new words appear dramatically, the dimension of vector, which represents the document is very large, we have to reduce the dimension by using various ways [3], [4]. And there are experimental results show that there is relation between words in a document indeed. Thus, this method draws back the performance of Information Retrieval(IR) system in some degree. This require us to optimize the model by another point of view. One of the solutions is dealing with the semantic connections between the words and the documents. We compare the documents in concept level. For the concept, 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 concept of Query Expansion and Ontology. Section 3 discusses the relevance computing. Section 4 presents the methods of query expansion. Finally, we conclude the paper with a summary and directions for future work in Section 5.

II. Related Works

A. Query Expansion

In an IR system, the user inputs his query sentence for the information and he hopes what he gets is what he wants. Normally, the terms in a query are not detailed enough to let the IR system understands what the user wants actually. Query expansion is one of the ways to solve this problem [5], [6].

Query expansion technology was brought forward in Ref. [6]. 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 query expansion. 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. [7]

B. Ontology

Ontology is explicit representations of a shared conceptualization, i.e., an abstract, simplied view of a shared domain of discourse. More formally, an ontology denes the vocabulary of a problem domain, and a set of constraints (axioms and rules) on how terms can be combined to model specic domains. An ontology is typically structured as a set of denitions 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.

Since the inception of the Semantic Web, in which ontology is the principal resource for integrating and dealing with online information, a new set of standards have been proposed. OWL is one such standard belonging to a family of knowledge representation languages prepared for the Semantic Web. OWL has attained the status of World Wide Web Consortium (W3C) recommendation. From a technical point of view, OWL extends the Resource Description Framework (RDF) and RDF Schema (RDFS), allowing us to integrate a variety of applications using XML as interchange syntax.

978-1-4673-0024-7/10/$26.00 ©2012 IEEE

C. Conceptual representation and indexing

In traditional IR system, documents are indexed by a set of words. Due to the ambiguity and the limited expressiveness of single words, it is difficult to decide which words should be expanded according to the terms in the query. For example, in VSM search model, there are millions similarity between documents and queries, the task of measuring is tremendous.

One way of improving the quality of similarity search is Latent Semantic Indexing(LSI) [8]. 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 Ref. [9], 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.

D. WordNet

WordNet is a well-known example of a machine-readable dictionary(MRD). It is one of the most important MRDs available to researchers in the fields of text analysis, computational linguistics, and many other related areas. It is an electronic lexical database designed by use of psycholinguistic and computational theories of human lexical memory. WordNet is composed of synonym sets(synsets), and each synset has an unique identifier(SynsetID). A synset is a list of word senses, and represents one constitutional lexicalized concept. It is unambiguous and carries exactly one meaning. 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 [10].

In this paper, we select one ontology from the WordNet to discuss our algorithm. So, it is named light ontology.

III. Relevance Computing

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

(1) Labeling the terms in the document. All documents will be represented by the terms, which should be belonged to one or more concepts. Thus, some terms in a document will be omitted if they cannot express the features of a document difference to others.
(2) Computing the weights of the terms to the concept. There are many terms in a concept. In a given domain, some terms are more important than others. We want to get a word list by the importance decrease order of the words to the concepts.
(3) Deciding the affiliation of the document to the domain. For a given document, it should be decided which domain the document belongs to. If two documents had same terms but different term orders, do they have same importance for a query or in a domain?

In the following, we construct the new approach to deal with the relevance of words and concepts, named W-C model. In this model, we concept the relevance of words to concepts. Further more, we construct the approach to consider the relevance between concepts and documents, named C-D model. Two model decide the relevance between the words, concepts, documents and queries.

A. Labeling Document

In our discussion, the first task is label the terms in a document. The term discussed here is the word or the phrase. The standard of labeling is WordNet, which has been introduced in the former. For the ontology and the document, we can assume the facts:

(1) An ontology is a very large set, and there are several hundreds of concepts, and there are many terms belong to each of the concepts. For a given term, it maybe belong to more than one concept.
(2) For a document, which is discussed for a given topic, it is impossible to include all of the terms in a special ontology.
(3) Assuming the word or phrase d belongs to a document $D(d \in D)$, d may be classified to concept C_1 or C_2 according to the term-list of the concepts. Which concept should be selected for the term d ? In most time, d is classified to C_1 or C_2 or both is decided by its neighbor terms set dN. In order to keep the document discussing the given topic smoothly, d should be classified to the same concept just as the most of its neighbor terms do.

We select the concepts in WordNet as the working level. The word-concept(W-C) model can be described just as follows:

(1) For each concept C_s, we construct a matrix U_sC, as,

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

Where u_sc_{ij} is the times which word d_i and d_j appear synchronously in a paragraph. u_sc_{ii} is the times which word d_i appear in a paragraph by oneself.
(2) Scanning the document D from the first word to the end, we label the words to the different concepts. If a word is belong to two or more concepts, labeling it to each of the concepts. For all of the documents, summing the times, of which word d_i and d_j appear synchronously, and replacing the value of element u_sc_{ij} in matrix U_sC.
(3) Dealing with the matrix U_sC. If the column i is all zero, it means the word d_i never appear in document D. We delete the column i and row i of this matrix.

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

The document D may have relevance with the concepts $C_{j1}, C_{j2}, \ldots, C_{jk}, ji \in [1, n]$. We denote the relevance by matrix $U_s C_p, p \in [1, n]$. In the following distribution , we write the matrix $U_s C_p, p \in [1, n]$ as Q for convenience.

B. Computing Relevance

We considered the matrix Q, whose elements $d_{ij}, (i, j \in [1, n])$ responded to the times of word pair d_i-d_j appeared in the same paragraph at the same time. 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 document D. Normalizing the matrix Q, we explain it as:

We have a set of words, $D = \{d_1, d_2, \cdots, d_n\}$, 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 $D = \{d_1, d_2, \cdots, d_n\}$ 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,

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

For the Markov chains, we can get

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

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

$$Q^{(n+m)} = Q^{(n)} \cdot Q^{(m)};$$
$$Q^{(2)} = Q^{(1)} \cdot Q^{(1)} = Q \cdot Q = Q^2;$$
$$Q^{(n)} = Q^{(n-1+1)} = Q^{(n-1)} \cdot Q^{(1)} = Q^{n-1} \cdot Q = Q^n.$$

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.

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 keywords on document D can be defined as a limiting solution of the iterative process:

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

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

C. Deciding affiliation

In this section, we will discuss the relevance of the document with the given topic. And we construct the concept-document(C-D) model to compute it.

Assuming the relevance of two documents D_1 and D_2 with the concept C_1 have been computed, we should decide which document will be list in first when the user's query is put forward?

The graph G discussed before is consisted of the nodes and the links. The node represents the words and the link represents the relation, which two words appear in the same paragraph. 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 relation between the words as,

Define 1: Assuming w_i and w_j are the nodes of graph G. If w_i does not connected to w_j directly, there is a path from w_i to w_j. We define 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.

The sum of the distance means the degree of words representing the concept or ontology. The more the sum , the much irrelevance of the words to the concept or the ontology.

Assuming the matrix Q_1 and Q_2 related to document D_1 and D_2 respectively, we set matrix $P_1 = Q_1$ and $P_2 = Q_2$, and $p_{ij} = 0, \forall i, j$. Computing the distance just as fellows,

(1) Computing the elements of P. According to the $Define1$ and $Define2$, the elements of P can be calculated.
(2) Computing the Average Variance of data series $\{p_{ij}, i \in [1, n] \wedge j \in [i, n].\}$
(3) Sum the Average Variance value

$$V = \sum_{i,j} p_{ij}, i \in [1, n] \wedge j \in [i, n].$$

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

IV. Methods of Query Expansion

When the user input the query to the search engine, the most important thing is to know what the user want to get exactly. In normal, the query sentence is not detail enough to be used to feedback the satisfactory results to the user. Query expansion can help to solve this problem. Ontologies play a key role in query expansion research. By using ontologies, we

can enrich the implication of query and to enhance the search capabilities of existing web searching systems.

By using the words in query sentence only, it is difficult to expand the words without any other help, such as the domain information, surfing history or log records. In order to solve the problem, the user is requested to register for the personalized service. The personal information, such as login name, interesting domains, professions and hobbies is included. The data are used to construct the Personal Information Profile(PIP). After the IR system feedback the results to the user, s/he looks through the results and estimate them. The IR system refine the PIP according to the estimation. The steps are,

(1) According the domains or fields name which the user represents in the register table, we set the weight of concepts in the domains or fields with 0.

(2) The user login the IR system and input the query, which means the user wishes to attend the development plan and wish to refine the feedback results according the PIP. We split the query to some words and mark 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.

(3) 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 the list that the words order by the relevance, which are computed in W-C model. We named it as,

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

   The second one is that the words order by the appearance 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,

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

   (c) Setting the threshold, and selecting the first P words. According to the relevance of words in P to the documents, those documents will be feedback to the user.

(4) The user reviews the results, and he presents his owner opinion for the retrieval course. The opinion will be used to refine the parameter $\alpha \in (0, 1)$ in the formula.

V. Conclusion

The paper discusses technologies of ontology, the method to present the document by using the ontology. We select part of the concepts in the WordNet to construct the Word-Concept model(W-C model). Computing the relevance between the words, we get a word list to describe the relation of words to the given concept. We construct the concept-document model(C-D model) by computing distance between the concept to the document. For the query expansion, the the Personal Information Profile(PIP) of user is built. 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 Scientific Research Funded by Xi'an University of Finance and Economics, P.R.China. (Program No.11XCK13).

References

[1] Z. Y.-M. L. C.-H. H. Y.-M. Tasi, C.-S.a, "Applying vsm and lcs to develop an integrated text retrieval mechanism," *Expert Systems with Applications*, vol. 39, no. 4, pp. 3974–3982, 2012.

[2] P. Cunningham, "A taxonomy of similarity mechanisms for case-based reasoning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 11, pp. 1532–1543, 2009.

[3] R. P. S.-H. Manning, C.D., "Introduction to information retrieval. cambridge university press," *Introduction to Information Retrieval*, 2008, cited By (since 1996) 1181.

[4] L. G. Gupta, V., "A survey of text mining techniques and applications," *Journal of Emerging Technologies in Web Intelligence*, vol. 1, no. 1, pp. 60–76, 2009.

[5] C. K.-S. Kim, M.-C., "A comparison of collocation-based similarity measures in query expansion," *Information Processing and Management*, vol. 35, no. 1, pp. 19–30, 1999.

[6] E. N. Efthimiadis, "Query expansion," *Annual Review of Information Science and Technology*, vol. 31, pp. 121–187, 1996.

[7] S. Cronen-townsend, Y. Zhou, and W. B. Croft, "A framework for selective query expansion," in *In Proceedings of Thirteenth International Conference on Information and Knowledge Management*. Press, 2004, pp. 236–237.

[8] S. T. Dumais, G. W. Furnas, T. K. Landauer, S. Deerwester, and R. Harshman, "Using latent semantic analysis to improve access to textual information," in *Proceedings of the SIGCHI conference on Human factors in computing systems*, ser. CHI '88. New York, NY, USA: ACM, 1988, pp. 281–285. [Online]. Available: http://doi.acm.org/10.1145/57167.57214

[9] A. Kayed and R. M. Colomb, "Using ontologies to index conceptual structures for tendering automation," in *Proceedings of the 13th Australasian database conference - Volume 5*, ser. ADC '02. Darlinghurst, Australia, Australia: Australian Computer Society, Inc., 2002, pp. 95–101. [Online]. Available: http://dl.acm.org/citation.cfm?id=563906.563917

[10] M. Dragoni, C. da Costa Pereira, and A. G. Tettamanzi, "A conceptual representation of documents and queries for information retrieval systems by using light ontologies," *Expert Systems with Applications*, vol. 10.1016/j.eswa.2012.01.188, no. 0, pp. –, 2012. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0957417412002163