IDSIS: INTELLIGENT DOCUMENT SEMANTIC INDEXING SYSTEM[21]

Zhongzhi Shi Bin Wu Qing He Xiujun Gong Shaohui Liu Yi Zheng
Shizz@ics.ict.ac.cn
Key Laboratory of Intelligent Information Processing ,
Institute of Computing Technology ,Chinese Academy of Sciences

Abstract: With rapid growth of the Internet, how to get information from this huge information space becomes an even important problem. In this paper, An Intelligence Document Semantic Indexing System : IDSIS is proposed. Some new technologies are integrated in IDSIS to obtain good performance. IDSIS is composed of four key procedures. A parallel, distributed and configurable Spider is used for information gathering; A multi-hierarchy document classification approach combining the information gain initially processes gathered web documents; A swarm intelligence based document clustering method is used for information organization; A concept-based retrieval interface is applied for user interactive retrieval. IDSIS is a concept-associated document semantic indexing system for information retrieval on Internet.

Keywords: document classification, document clustering, semantic indexing, concept space

1. INTRODUCTION

Large quantities of textual data available for example on the Internet pose a continuing challenge to applications that help users in making sense of the data. We describe here a new approach to automatic text classifying, clustering, indexing and topic retrieval. It is designed to overcome a fundamental problem that plagues existing retrieval techniques that try to

[21] This research supported by NSFC grant 60073019, 90104021, 60173017 and NSFB grant 4011003.

The original version of this chapter was revised: The copyright line was incorrect. This has been corrected. The Erratum to this chapter is available at DOI: 10.1007/978-0-387-35602-0_35

M. A. Musen et al. (eds.), *Intelligent Information Processing*

match words of queries with words of documents. The problem is that users want to retrieve on the basis of conceptual content, and individual words provide unreliable evidence about the conceptual topic or meaning of a document. There are usually many ways to express a given concept. The literal terms in a user's query may not match those of a relevant document. In addition, most words have multiple meanings, so terms in a user's query will literally match terms in documents that are not of interest to the user. The proposed approach tries to overcome the deficiencies of term matching retrieval by treating the unreliability of observed term document association data as a statistical problem.

2. ARCHITECHURE OF IDSIS

An automatic document collector is one of the fundamental components of IDSIS. The main task performed by this component is the automatic gathering of Web documents, which are usually stored locally for further processing. A parallel and distributed spider is employed in this system [3]. The documents collected by spider are processed by text analyzer including parsing and extracting the informative words for classification, concept clustering, semantic index generating and so on. Text classifier learned knowledge from pre-labeled documents and then classified the newly documents. An innovative and multi-layer classification model is used [4]. With the increase of Internet documents, the volumes of some directors become larger. To decrease the search space for retrieval by directory, we provide a document Clustering algorithm based on Swarm Intelligence and k-Means: CSIM [6]. CSIM gives the class description for each cluster. With the assistance of concept space, the system could help the searchers to find most relevant terms and retrieve most relevant documents by using these terms. In our system, we incorporate these two methods to generate concept space. We use Chen's[2] method to generate the concept space of specific domain and use HowNet [7] to generate the concept of general domain. By building the concept semantic space, IDSIS can broad or locate what they most want. Users can not only browse all documents belong some directory, but also gives the association concepts with the class concept in the directory way. IDSIS will return its association concepts and documents related the few words provided by users. Meanwhile, users may also filter some documents by setting the documents' class. Limited by the paper length, we only briefly introduce two parts of IDSIS. They are text classification and clustering.

3. TEXT CLASSIFICATION: INFORMATION INITIAL PROCESSING

After the web pages are crawled by Spiders, they will be classified automatically. To calculate term weight, the TF.IDF approach[5] consider two factors: TF (term frequency) and IDF (inverse document frequency).We introduces the concept of Information Gain from Information Theory, i.e., the document collection D is regarded as an information source according with some probability distribution, and the amount of information provided by one term (the importance of the term) in text classification can be obtained by considering the information gain between the information entropy of document collection and the conditional entropy of the term, then we can combine this amount of the information into the formula of calculating term weight.

The amount of information of term T_k can be calculated using the following formula: $IG_k=H(D)-H(D/T_k)$, Where H(D) is the entropy of the document collection D: $H(D) = -\sum_{d_i \in D} P(d_i) \times \log_2 P(d_i)$, $H(D/T_k)$ is the conditional entropy of term T_k: $H(D \mid T_k) = -\sum_{d_i \in D} P(d_i \mid T_k) \times \log_2 P(d_i \mid T_k)$, $P(d_i)$ is the conditional probability of the document d_i: $P(d_i) = \frac{wordfreq\ (d_i)}{\sum_i wordfreq\ (d_i)}$, where wordfreq($d_i$) is the sum of all the term frequencies in d_i. Then TF.IDF formula can be revised as follows:

$$W_{ik} = \frac{tf_{ik} \times \log(\ N\ /\ n_k\ + 0.01) \times IG_k}{\sqrt{\sum_{k=1}^{m} (tf_{ik})^2 \times [\log(\ N\ /\ n_k\ + 0.01) \times IG_k]^2}} \quad (1)$$

We also propose an approach of multi-hierarchy text classification based on VSM, i.e., all classes are organized as a tree according to some given hierarchical relations. The basic insight supporting our approach is that classes that are attached to the same node have a lot more in common with each other than else classes, so the models of these classes will be based on a small set of features. In the test phase, we collect 21,430 documents downloaded from some famous web sits in China such as SINA, FM365 as training set. The corporation collects 11062 documents as testing set. There are 34 classes. The average precision is 84.20%.

4. TEXT CLUSTERING: SEMANTIC INFORMATION ORGANIZATION

We propose a document Clustering algorithm based on Swarm Intelligence and k-Means: CSIM in this paper. Swarm Intelligence is defined as any attempt to design algorithms or distributed problem-solving devices inspired by the collective behavior of the social insect colonies and other animal societies [1]. The main idea of CSIM is that: firstly, documents are denoted

by Vector Space Model, and each vector regarded as a data object is randomly projected onto a plane. Secondly, each simple agent(ant) perceive the swarm similarity of current object within the local region by formula (2), and compute the picking-up or dropping probability, and act according to this probability. Clusters will visually form on the ant-work plane through simple agents' collective actions. Then, clusters results are collected from the plane by a recursive algorithm. Finally, an iterative partitioning phase is employed to further optimize the results. A formula of measuring the swarm similarity between a pair of documents is showed as formula (2).

$$f(d_i) = \sum_{O_j \in Neigh(r)} \left[1 - \frac{10*(1-sim(d_i,d_j))}{\alpha} \right] \quad (2)$$

where $Neigh(r)$ denotes the local region, it is usually a rounded area with a radius r. $d(o_i,o_j)$ denotes the distance of document vector o_i with o_j in the space of attributes. It is usually cosine distance. The parameter α is defined as swarm similarity coefficient. The functions to compute picking-up or dropping probability are two lines with a slope k.

5. CONCLUSIONS

In this paper, an all-sided solution for information retrieval on Internet (IDSIS) is proposed. Concept search is a trend of information retrieval. IDSIS partially realized a concept-associated search by efficient information organization. IDSIS covers all procedures of information retrieval on internet, i.e. information gathering, information organization and information querying.

6. REFERENCES

[1]. E.Bonabeau, , M.Dorigo, & G.Theraulaz, Swarm Intelligence: From Natural to Artificial Systems, Oxford Univ. Press, New York, 1999

[2]. H. Chen, K. J. Lynch, K. Basu, and D. T. Ng. Generating, integrating, and activating thesauri for concept-based document retrieval. IEEE EXPERT, Special Series on Artificial Intelligence in Text-based Information Systems, 8(2):25-34, April 1993.

[3]. Dong Mingkai, Tian Qijia, Shi Zhongzhi, Web Spider Based on Intelligent Agent. SCI2001, Orlando, P292-296, 2001.

[4]. Shaohui Liu, Mingkai Dong, Haijun Zhang, Rong Li, Zhongzhi Shi, An Approach of Multi-hierarchy Text Classification, 2001 International Conferences on Info-tech and Info-net PPOCEEDINGS,Conferences C:95-100

[5]. G. Salton, B. Buckley. Term-weighting Approaches in Automatic Text Retrieval. Information Processing and Management, 1998, 24(5) : 513-523

[6]. Wu Bin Zheng Yi Liu Shaohui Shi Zhongzhi,CSIM: A Document Clustering Algorithm Based On Swarm Intelligence, To be appeared In Proceedings of Congress on Evolutionary Computation,2002

[7]. Zhendong Dong, Qiang Dong, HowNet, http://www.keenage.com