International Journal of Advanced Computer Technology (IJACT)

ISSN:2319-7900

A CONCEPT DRIVEN SEMANTIC APPROACH FOR WEB DOCUMENT CLUSTERING

Poonam Churi,Department of computer Engneering,Datta Meghe college Of Engineering,Airoli,Navi Mumbai,India ;
Sujata Kolhe,Department of Information Technology,,Datta Meghe college Of Engineering,Airoli,Navi Mumbai,India

Abstract

In the age of increasing information availability, many techniques, such as document clustering, Web search result clustering and information visualization, have been developed to ease understanding of information for users. Organizing Web search results into clusters ease users' quick browsing through search results. However, most of Traditional clustering techniques do not help users directly understand key concepts and their semantic relationships in document corpora, which are critical for capturing their conceptual structures. These clustering techniques are least adequate as they don't suggest clusters with highly readable names. Therefore, we present a novel approach called 'Semantic Lingo' to identify the key concepts and automatically generate ontology based on these concepts for conceptualization of document.

Keywords

Document clustering, Web search Result, Relevant, Domain ontology.

1. Introduction

Extensive growth in information technology and its exponential use has generated myriad of data in the forms of text documents. In order to have fast retrieval from such extensive data and to facilitate effective browsing and searching it is vital to organize documents. Search engines are considered as the most common tool to retrieve information from the Internet. But the search results returned by search engines usually come with huge quantity and a long list. What's more, the search results will be very diverse when users' search term is not correct or ambiguous. It's time-consuming and arduous to find the most relevant search result. Search results clustering are an efficient method to make the search results easier to scan. Search results clustering works on snippets (a summary of the search results), which is different from document clustering [1] (working on long text).Traditional clustering algorithms do not consider the semantic relationships among the words so that can not accurately represent the meaning of documents. To overcome this problem semantic information from ontology such as Domain Ontology has been used to improve the Quality of Web search clustering. Our key goal is to improve our system by overcome various problems such as Synonym and polysemy, high dimensionality and assigning appropriate description for generated cluster. [1]

Traditional clustering algorithms are relying on the BOW (Bag of Words) approach and a disadvantage of the BOW is that it ignores the semantic relationship among words so that can not accurately represent the meaning of document. As the rapid growth of text documents, the textual data have become diversity of vocabulary, they are high dimensional. However, most of these methods only provide ways for users to easily access the information; they do not help users directly capture the key concepts and their relationships within the information. Key concepts are the representative concepts that best describe a document corpus. [1]. In this paper, we propose a novel approach called 'Semantic Lingo' to identify the key concepts and automatically generate corpus-related ontology's for conceptualizing document Conceptualization of document corpora here means representing document corpora with a set of concepts and their relationships, which can help users more easily understand what the documents are about and the semantic relationships contained therein. Semantic Lingo applies latent semantic analysis to identify key concepts, allocates documents based on these concepts, and utilizes WordNet as Domain Ontology [7] to automatically generate a corpus-related ontology. Finally, the documents are linked to the ontology through the key concepts [3]

2. Related Work

To help users better understand the structure of document corpora several clustering algorithms that extract meaningful labels for documents have been proposed. Zamir et al.[19] proposed a phrase-based document clustering approach based on suffix tree clustering. The first algorithm to take the approach based on frequent phrases shared by documents in the collection was put forward in 1998 and called Suffix Tree Clustering (STC) [2]. Laterin 2001, the SHOC (Semantic, Hierarchical, and Online Clustering) algorithm was introduced [5]. SHOC improves STC and is based on LSI and frequent phrases. Next in 2003, the Lingo algorithm [7] was devised. This algorithm is used by the Carrot2 web searcher and it is based on complete phrases and LSI with Singular Value Decomposition (SVD).Lingo is an improvement of SHOC and STC and (unlike most algorithms), tries first to discover descriptive names for the clusters and only then organizes the documents into appropriate clusters. NMF (also in 2003) is another example of these algorithms, it is based on the non-negative matrix factorization of the term-document matrix of the given document corpus was made available [9]. This algorithm surpasses the LSI and the spec-

ISSN:2319-7900

tral clustering methods in document clustering accuracy but does not care about cluster labels. Another approach was proposed by the Pair wise Constraints guided Non-negative Matrix Factorization (PCNMF) algorithm [9] in 2007. This algorithm transforms the document clustering problem from an un-supervised problem to a semi-supervised problem using must-link and cannot-link relations between documents.

Osinski et al. [4] proposed a concept-driven algorithm for clustering search results, the Lingo algorithm, which uses LSI (Latent Semantic Indexing) techniques to separate search results into meaningful groups. However, most of the above clustering methods are focused on separating documents into meaningful groups only; they do not consider building semantic relationships between these groups.

3. Proposed System

Figure 1 shows the overview of Semantic Lingo. First, a document corpus is preprocessed into term frequency files, in which each document is represented as a list of its term frequencies. In addition, common phrases are extracted using a suffix array algorithm [9]. Second, the inverted document frequency of each term is calculated and each term weight is computed by multiplying the term frequency and inverted document frequency. Inverted term-document files are generated for each term and the term-document matrix is constructed based on term weights. Third, with the extracted common phrases, we conduct key concept induction using LSA techniques. Fourth, with the list of key concepts, we utilize WordNet to inspect their synonyms and hyponyms. The documents are allocated based on each key concept and its synonyms and hyponyms. Fifth, using WordNet, hypernyms of each concept are detected and used to construct a corpus-related ontology. Sixth, documents are linked to the ontology through the key concepts. To capture the key concepts for a document corpus, as in [6] we use the vector space model (VSM) and singular value decomposition (SVD), the latter being the fundamental mathematical construct underlying the latent semantic analysis (LSA) technique. VSM is a method of information retrieval that uses linear-algebra operations to compare textual data, associating a single multidimensional vector with each document in a collection, and each component of that vector reflects a particular keyword or term related to the document. LSA aims to represent the input collection using abstract terms found in the documents rather than the literal terms appearing in them, by approximating the original term-document matrix using a limited number of orthogonal factors. These factors represent a set of abstract terms, each conveying some idea common to a subset of the document corpora. In Semantic Lingo, these terms are used as candidate concepts for representing the document corpus. In Semantic Lingo, phrases that appear at least a certain number of times in the document corpus are also considered possibly then; documents are clustered and assigned to each group based on the key concepts as well as their synonyms and their hyponyms. Using Word- Net, hypernyms of these concepts are detected and used to generate a corpus-related ontology automatically. Documents are linked to the ontology through the key concepts meaningful to users, as they are often collocations or proper names.

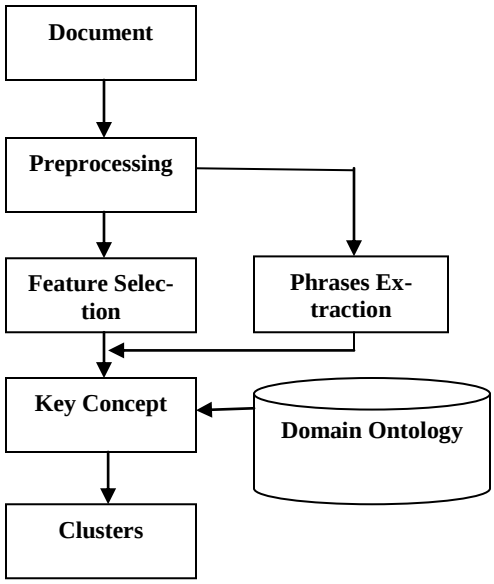

Figure.1 Overview of System.

3.1 Preprocessing and common phrases extraction

This component converts each of the search results (as snippets) into a sequence of words, phrases, strings or general attributes or characteristics, which are then used by the clustering algorithm. There are a number of tasks performed on the search results, including: removing special characters and accents, the conversion of the string to lowercase, removing stop words, stemming of the words and the control of terms or concepts allowed by a vocabulary [6] To extract the common phrases from the document corpus, we use the suffix array algorithm proposed by Yang et al. [2]. To filter out the meaningless phrases, we define some rules. First, the phrases appear in the document corpus at least a specified number of times. Second, the phrases do not cross sentence boundaries. Third, if a phrase is contained in a longer phrase, the latter is regarded as more complete than the former. Fourth, if a phrase begins with a stop word or ends with a stop word, it is regarded as meaningless and discarded. Note that if a phrase contains a stop word in the middle of it, we do not discard it because a meaningful phrase can contain a stop word

ISSN:2319-7900

3.2 Feature Selection and Key concepts Creation

Feature selection done by VSM (vector space model) and LSI (latent semantic Indexing) methods. In the Vector Space Model (VSM) [9] every document in the collection is represented by a multidimensional vector. Each component of the vector reflects a particular key word or term connected with the given document. The value of each component depends on the degree of relationship between its associated term and the respective document. Many schemes for measuring the relationship, very often referred to as term weighting, have been proposed. In the following subsection we review the three most popular.

3.2.1Term weighting

Term weighting is a process of calculating the degree of relationship (or association) between a term and a document. As the Vector Space Model requires that the relationship be described by a single numerical value, let aij represent the degree of relationship between term i and document j.In the simplest case the association is binary: aij=1 when key word i occurs in document j, aij=0 otherwise. The binary weighting informs about the fact that a term is somehow related to a document but carries no information on the strength of the relationship. A more advanced term weighting scheme is the term frequency. In this scheme aij=tfij where tfij denotes how many times term i occurs in document j. The tf-idf (term frequency inverse document frequency) scheme aims at balancing the local and the global term occurrences in the documents. In this scheme aij=tfij· log (N/dfi) where tfij is the term frequency, dfi denotes the number of documents in which term i appears, and N represents the total number of documents in the collection. The log (N/dfi), which is very often referred to as the idf (inverse document frequency) factor, accounts for the global weighting of term i. Indeed, when a term appears in all documents in the collection, dfi=N and thus the balanced term weight is 0, indicating that the term is useless as a document discriminator. [7, 9]

3.2.2 Query matching

In the Vector Space Model, a user query is represented by a vector in the column space of the term-document matrix. [7] This means that the query can be treated as a pseudo-document that is built solely of the query terms. Therefore, in the process of query matching, documents must be selected whose vectors are geometrically closest to the query vector. A common measure of similarity between two vectors is the cosine of the angle between them. In a t×d term-document matrix A, the cosine between document vector aj and the query vector q can be computed according to the formula:

$$\cos\theta_j = \frac{a_j^T q}{\|a_j\|\|q\|} = \frac{\sum_{i=1}^{t} a_{ij} q_i}{\sqrt{\sum_{i=1}^{t} a_{ij}^2}\sqrt{\sum_{i=1}^{t} q_i^2}} \quad (1)$$

3.2.3 LSI

The fundamental mathematical construct underlying the LSI is the Singular Value Decomposition [9] of the term-document matrix. The decomposition breaks a t×d matrix A into three
matrices U, Σ and V such that A=UΣVT U is the t×t orthogonal matrix whose column vectors are called the left singular vectors of A, V is the d×d orthogonal matrix whose column vectors are termed the right singular vectors of A, and Σ is the t×d diagonal matrix having the singular values of A ordered decreasingly (σ1 ≥ σ2 ≥ ... ≥ σmin (td)) along its diagonal. The rank rA of the matrix A is equal to the number of its non-zero singular values. The first rA columns of U form an orthogonal Using the Singular Value Decomposition a k-rank approximation AK of matrix A can be calculated that is closest16 to the original matrix for a given value of k. The AK matrix can be obtained from the SVD-derived matrices by setting all but the k largest singular values in the Σ matrix to 0. Thus, Ak=UkΣkVk T, where Uk is the k×t matrix whose columns are first k columns of U, Vk is the d×k matrix whose columns are the first k columns of U, and Σk is the k×k matrix whose diagonal elements are the klargest singular values of A.[10]

3.3 Cluster Generation Using Domain Ontology

To allocate the documents, we match them against the related concepts. This process is similar to extracting related documents for a query in information retrieval models. Semantic Lingo uses the key concepts as queries. Similar to the ideally used method query expansion, we utilize Word-Net
[6] to inspect synonyms and hyponyms of the key concepts among the frequent terms in the document corpus. The hyponyms within five levels of each concept are examined by Semantic Lingo. [3] Documents are allocated to each key concept based on the cosine similarity between each document and the set including the key concept and its synonyms and hyponyms. For each concept, if the cosine similarity between a document and the concept exceeds a predefined threshold, the document is allocated to the corresponding group represented by the concept. This assignment method naturally creates overlapping groups and well handles crosstopic documents. Note that not only key concepts, but also their synonyms and hyponyms are used to generate ontologies.The use of synonyms and hyponyms can help users easily visualize documents related to the concepts similar to or more specific than the key concepts. To help users capture

ISSN:2319-7900

the conceptual structure more intuitively, we flatten the generated ontology by amalgamating with their parent nodes the 'only-child' nodes that do not belong to the frequent terms.

4. Conclusion

In this paper, we proposed a novel method, Semantic Lingo which identifies key concepts and automatically generates Ontology's for users to conceptualize document corpora. Based on the vector space model, LSA techniques are used to identify the meaningful key concepts. The documents are allocated to these concepts using cosine similarities. By inspecting the hypernyms of these concepts, ontology's are automatically generated and documents are linked to the generated ontology's through the concepts. The experimental results show that Semantic Lingo can extract key concepts from document corpora with a high precision. In this paper, we have presented a new framework in which our system provides solution to Problems related to the document clustering. A modified WordNet based algorithm semantic similarity measure is proposed for word sense disambiguation. We will try to solve problems in text clustering such as polysemous and synonyms words, high dimensionality and properly assign documents to clusters. It gives implicit and explicit relationship between documents. We propose a novel method, Semantic Lingo which identify concepts and automatically generate ontology's for users to conceptualize document corpora. Based on the vector space model, LSA techniques are used to identify the meaningful key concepts. The documents are allocated to these concepts using cosine similarities .This method naturally creates overlapping groups and well handles cross topic documents. By inspecting the hyponyms of these concepts, ontology's are automatically generated and documents are linked to the generated ontology's through the concepts. As we compare different web search clustering algorithm like STC and SHOC does not reduce the high dimension of the text documents, hence its complexity is quite high for large text databases, which ignores the semantic and lexical relationships between words, so we proposed new algorithm called 'Lingo Semantic' for clustering

References

[1] Web Document Clustering: A Feasibility Demonstration" Oren Zamir and Oren Etzioni Department of Computer Science and Engineering University of Washington Seattle, WA 98195-2350 U.S.A.

[2] The Suffix Tree Document Model Revisited Sven Meyer zu Eissen (Paderborn University, Germany smze@upb.de) Benno Stein (Bauhaus University Weimar, Germany benno.stein@medien.uniweimar.

[3] Martin Potthast (Paderborn University, Germany beebop@upb.de A Search Result Clustering Method using informatively Named Entities" Hiroyuki Toda NTT Cyber

[4] 94-5/05/0011..$5.00 S.Osinski and D. Weiss. A concept-driven algorithm for clustering search results. 20(3):48–54, 2005.] [5] M. Steinbach, et al., "A comparison of document clustering techniques," in KDD workshop on text mining, Boston, MA, USA., 2000, pp. 1-20

[6] Tingting Wei, Yonghe Lu, Huiyou Chnag, Qiang Zhou, Xianyu Bao ," Sematic approach for text clustering using WordNet and lexical chains", Expert Systems with Applications,Volume 42, Issue 4, March 2015, Pages 2264–2275.

[7] Stanislaw Osiríski and Dawid Weiss. Conceptual clustering using Lingo algorithm: Evaluation on Open Directory Project data. Submitted to Intelligent Information Systems Conference 2004, Zakopane, Poland, 2003

[8] Yang, X.S., Deb, S.: Cuckoo search via Levy flights. In: Proc. of the World Congress on Nature and Biologically Inspired Computing, India, pp. 210–214 (2009)

[9] Andrea Tagarelli, George Karypis," A segment-based approach to clustering multi-topic documents", Knowl INF Syst (2013) Springer-Verlag London Limited 2012.

[10] Moe Moe Zaw and Ei Ei Mon,Web Document Clustering Using Cuckoo Search Clustering Algorithm based on Levy Flight, International Journal of Innovation and Applied Studies ISSN 2028-9324 Vol. 4 No. 1 Sep. 2013

[11] S. E. Robertson and K. Sparck-Jones, "Relevance ighting of search terms," in *Document retrieval systems*, Ed: Taylor Graham Publishing, 1988, pp. 143-160.

[12] Mohammad Khabbaz, Keivan Kianmehr, and Reda Alhajj, Employing Structural and Textual Feature Extraction for Semistructured Document Classification, IEEE Transactions on Systems, Man, and Cybernatics-Part C: Applications and Reviews, Vol. 42, No. 6, November 2012.

[13] Jian-Wen Zhao; Shen-Ming Gu; Ling He "A novel approach to clustering access patterns in e-learning environment", Education Technology and Computer (ICETC), 2010 2nd International Conference on page(s): V1-393 - V1-397 Volume: 1, 22-24 June 2010

[14] Lingras, P.; Rathinavel, K. "Recursive meta-clustering in a granular network" Intelligent Systems Design and Applications (ISDA), 2012 12th International Conference on, on page(s): 770 – 775

[15] Voges, K.E.; Pope, N.K.L. "Rough Clustering Using an Evolutionary Algorithm", System Science (HICSS), 2012 45th Hawaii International Conference on, On page(s): 1138 - 1145