ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Keyword Extraction and Clustering for Documentation in Conversation

Manasa.N, SMT. Kavitha. G

PG Scholar, Department of Computer Science & Engineering, UBDT College of Engineering, Davangere, India,

Assistant Professor, Department of Computer Science & Engineering, UBDT College of Engineering, Davangere, India

ABSTRACT: Humans are surrounded by an unprecedented wealth of information, available as documents, databases, or multimedia resources. Access to this information is conditioned by the availability of suitable search engines, but even when these are available, users often do not initiate a search, because their current activity does not allow them to do so, or because they are not aware that relevant information is available. We adapt the perspective of just-in-time retrieval, which answers this shortcoming by spontaneously recommending documents that are related to users' current activities.

KEYWORDS: keyword, extraction, clustering.

I. Introduction

The task of recommending documents to knowledge workers differs from the task of recommending products to consumers. Collaborative approaches [1, 2, 3], as applied to books, videos and the like, attempt to communicate patterns of shared taste or interest among the buying habits of individual shoppers to augment conventional search results. There are well-known problems with these approaches, e.g., when consumers temporarily shop for their children, but their effectiveness has been established in practice at ecommerce sites such as Amazon.

It turns out that subtle variations in search context can undermine the effectiveness of collaborative filtering. For example, a lawyer might research one side of a case today, and tomorrow want to argue the other side of a similar case. This is rather like the 'shopping for children' example, in which a consumer's tastes and interests appear to change capriciously, from the system's point of view. Also, lawyers are reluctant to share their search history with others for a variety of reasons, ranging from confidentiality to competitive advantage.

Content-based recommendation systems analyze item descriptions to identify items that are of particular interest to the user. Because the details of recommendation systems differ based on the representation of items, this chapter first discusses alternative item representations. [2] Next, recommendation algorithms suited for each representation are discussed. The chapter concludes with a discussion of variants of the approaches, the strengths and weaknesses of content-based recommendation systems, and directions for future research and development.

The role of images content and metadata: In general, similar images often incur similar privacy preferences, especially when people appear in the images. Analyzing the visual content may not be sufficient to capture users' privacy preferences.[7] Tags and other metadata are indicative of the social context of the image, including where it was taken and why and also provide a synthetic description of images, complementing the information obtained from visual content analysis.

II. RELATED WORK

Khalid Al-Kofahi et.al [1] has proposed the task of recommending documents to knowledge workers differs from the task of recommending products to consumers. Variations in search context can undermine the effectiveness of collaborative approaches, while many knowledge workers function in an environment in which the open sharing of information may be impossible or undesirable. There is also the ‘cold start’ problem of how to bootstrap a recommendation system in the absence of any usage statistics. he describe a system called ResultsPlus, which uses a blend of information retrieval and machine learning technologies to recommend secondary materials to attorneys engaged in primary law research. Rankings of recommended material are subsequently enhanced by incorporating historical user behavior and document usage data.

Michael J. et.al proposes a [2] systems that recommend an item to a user based upon a description of the item and a profile of the user's interests. Content-based recommendation systems may be used in a variety of domains ranging from recommending web pages, news articles, restaurants, television programs, and items for sale. Although the details of various systems differ, content-based recommendation systems share in common a means for describing the items that may be recommended, a means for creating a profile of the user that describes the types of items the user likes, and a means of comparing items to the user profile to determine what to recommend. The profile is often created and updated automatically in response to feedback on the desirability of items that have been presented to the user.

The task of recommending content to professionals [3] (such as attorneys or brokers)differs greatly from the task of recommending news to casual readers. A casual reader may be satisfied with a couple of good recommendations, whereas an attorney will demand precise and comprehensive recommendations from various content sources when conducting legal research. Legal documents are intrinsically complex and multi-topical,contain carefully crafted, professional, domain specific language, and possess a broad and unevenly distributed coverage of issues. Consequently, a high quality content recommendation system for legal documents requires the ability to detect significant topics from a document and recommend high quality content accordingly.

Sangeetha. J et.al [4] To provide security for the information, automated annotation of images are introduced which aims to create the meta data information about the images by using the novel approach called Semantic annotated Markovian Semantic Indexing(SMSI) for retrieving the images. The proposed system automatically annotates the images using hidden Markov model and features are extracted by using color histogram and Scale-invariant feature transform (or SIFT) descriptor method. After annotating these images, semantic retrieval of images can be done by using Natural Language processing tool namely Word Net for measuring semantic similarity of annotated images in the database. Experimental result provides better retrieval performance when compare with the existing system.

Using social media we are able to communicate with lot of people. Facebook is most popular example of social media which enable us to communicate with lot of people. In which peoples have opportunities to meet new peoples, friends and communicate with each other.[5] In this paper author concentrated on Social media, content sharing sites, Privacy, Meta data. We propose a two-level [6] framework which according to the user's available history on the site, determines the best available privacy policy for the user's images being uploaded. Our solution relies on an image classification framework for image categories which may be associated with similar policies, and on a policy prediction algorithm to automatically generate a policy for each newly uploaded image, also according to users' social features.

III. PROPOSED SYSTEM

Here propose an efficient way for document recommendation system for user using the conversational data. Set of conversational data is given as input. These conversational data is partitioned into n clusters. Clusters contain several numbers of keywords including unwanted words. Using dictionary only relevant and meaningful keywords are extracted.

Keywords are ranked based on their weights. By choosing highest ranked keyword document recommendation process will achieved.

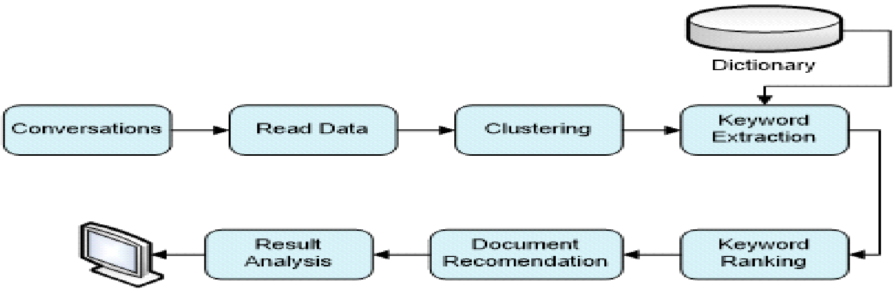

Figure1: Architecture of proposed system

A. Keyword Extraction

Term Frequency - Inverse Document Frequency or simply TF-IDF, weights a given term to determine how well the term describes an individual document within a corpus. It does this by weighting the term positively for the number of times the term occurs within the specific document, while also weighting the term negatively relative to the number of documents which contain the term. Consider term t and document d ∈ D, where t appears in n of N documents in D. The TF-IDF function is of the form:

T F IDF (t, d, n, N) = T F (t, d) × IDF (n, N) There are many possible TF and IDF functions.

Practically, nearly any function could be used for the TF and IDF. Regularly-used functions include:

$$\mathrm{TF}(t, d) = \begin{cases} 1 \text{ if } t \in d \\ 0 \text{ else} \end{cases}$$

$$\mathrm{TF}(t, d) = \sum_{\mathrm{word} \in d} \begin{cases} 1 \text{ if word} = t \\ 0 \text{ else} \end{cases}$$

When the TF-IDF function is run against all terms in all documents in the document corpus, the words can be ranked by their scores. A higher TF-IDF score indicates that a word is both important to the document, as well as relatively uncommon across the document corpus. This is often interpreted to mean that the word is significant to the document, and could be used to accurately summarize the document. TF-IDF provides a good heuristic for determining likely candidate keywords, and it (as well as various modifications of it) have been shown to be effective after several decades of research. Several different methods of keyword extraction have been developed since TF-IDF was first published in 1972, and many of these newer methods still rely on some of the same theoretic backing as TF-IDF. Due to its effectiveness and simplicity, it remains in common use today.

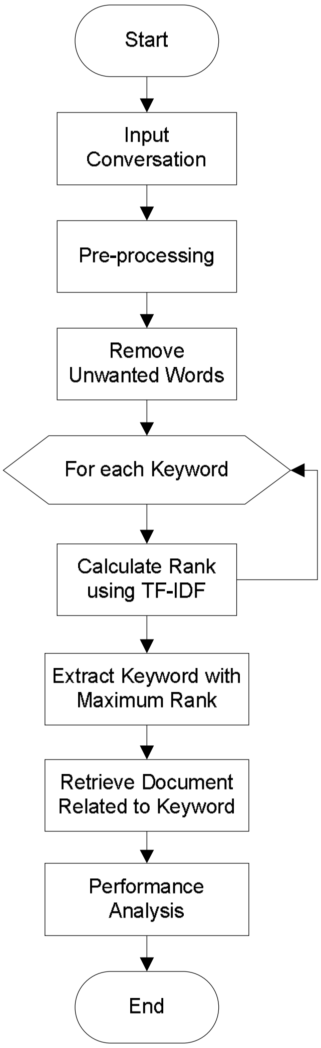

Figure 2: Flowchart of Proposed architecture

B. Clustering

The diverse set of extracted keywords is considered to represent the possible information needs of the participants to a conversation, in terms of the notions and topics that are mentioned in the conversation. To maintain the diversity of topics embodied in the keyword set, and to reduce the noisy effect of each information need on the others, this set must be split into several topically-disjoint subsets. Each subset corresponds then to an implicit query that will be sent to a document retrieval system. These subsets are obtained by clustering topically-similar keywords, as follows. Clusters of keywords are built by ranking keywords for each main topic of the fragment.

C. Keyword to Document Recommandation

As a first idea, one implicit query can be prepared for each conversation fragment by using as a query all keywords selected by the diverse keyword extraction technique. However, to improve the retrieval results, multiple implicit queries can be formulated for each conversation fragment, with the keywords of each cluster from the previous section, ordered as above (because the search engine used in our system is not sensitive to word order in queries). In experiments with only one implicit query per conversation fragment, the document results corresponding to each conversation fragment were prepared by selecting the first document retrieval results of the implicit query. The recommendation lists were prepared by selecting the first document retrieval results of each implicit query and then ranking documents based on the topical similarity of their corresponding queries to the conversation fragment.

IV. RESULT AND DISCUSSION

The experiment is tested on 50 queries taken from twitter, the conversations are given in the form of text files. By using language model and clustering, keywords are extracted. Then, each keyword is ranked based on its frequency in the database. Finally most ranked keyword is chosen as keyword for document recommendation. The analysis table is shown in table 1.

	Number of Queries	Keyword Relevancy	irrelevant
Existing System	50	80%	20%
Proposed Method	50	88%	12%

Table 1: Result Analysis

V. CONCLUSION

Our current goals are to process also explicit queries, and to rank document results with the objective of maximizing the coverage of all the information needs, while minimizing redundancy in a shortlist of documents. In our proposed system. We have considered a retrieval systems intended for conversational environments, in which they recommend to users documents that are relevant to their information needs. Enforcing both relevance and diversity brings an effective improvement to keyword extraction and document retrieval.

REFERENCES

[1] Khalid Al-Kofahi, Peter Jackson, Mike Dahn*, Charles Elberti, William Keenan, John Duprey.A "Document Recommendation System Blending Retrieval and Categorization Technologies".
[2] Michael J. Pazzani and "Daniel Billsus, Content-based Recommendation Systems "..
[3] Qiang Lu and Jack G. Conrad, "BringingOrdertoLegalDocuments AnIssue-basedRecommendationSystemviaClusterAssociation".Thomson Reuters Corporate Research & Development when this work was conducted.
[4] Sangeetha. J 1, Kavitha. R ," An Improved Privacy Policy Inference over the Socially Shared Images with Automated Annotation Process ", / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 3166-3169.
[5] Aishwarya Singh, Bhavesh Mandalkar, Sushmita Singh , Prof. Yogesh Pawar, "A Survey on User-Uploaded Images Privacy Policy Prediction Using Classification and Policy Mining",International Journal of Innovative Research in Computer and Communication Engineering .Vol. 3, Issue 9, September 2015.
[6] X. Liu, X. Zhou, Z. Fu, F. Wei, and M. Zhou, "Exacting social events for tweets using a factor graph," in Proc. AAAI Conf. Artif. Intell., 2012, pp. 1692–1698.
[7] A. Cui, M. Zhang, Y. Liu, S. Ma, and K. Zhang, "Discover breaking events with popular hashtags in twitter," in Proc. 21st ACM Int. Conf. Inf. Knowl. Manage., 2012, pp. 1794–1798.
[8] A. Ritter, Mausam, O. Etzioni, and S. Clark, "Open domain event extraction from twitter," in Proc. 18th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, 2012, pp. 1104–1112.
[9] X. Meng, F. Wei, X. Liu, M. Zhou, S. Li, and H. Wang, "Entitycentric topic-oriented opinion summarization in twitter," in Proc. 18th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, 2012, pp. 379–387.
[10] X. Wang, A. McCallum, and X. Wei, "Topical n-grams: Phrase and topic discovery, with an application to information retrieval," in Proc. IEEE 7th Int. Conf. Data Mining, 2007, pp. 697–702.
[11] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in Proc. 13th Conf. Comput. Natural Language Learn., 2009, pp. 147–155.