Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2016, 3(4): 12-17

Review Article ISSN: 2394 - 658X

Keyword Extraction Mechanism for Efficient Document Recommendation: A Review

Shreya S Surjuse and SC Dharmadhikari

Department of Information Technology, Pune Institute of Computer Technology, Pune, India
sshreya.456@gmail.com

ABSTRACT

Document recommendation system is useful to find the right results immediately without adding filters manually or clicking through multiple navigation menus. This paper first surveys various just in time retrieval agent and then reports several term weighting scheme. Just in time retrieval agent proactively provides valuable information without any explicit search. Term weighting enriches the retrieval efficiency. Previously recommendation was based on the word frequency, but to make the recommendation relevant, coverage of main topic becomes important. The comparative study among them will help in selecting the appropriate method.

Key words: Keyword extraction, Term weighting, Document recommendation

---

INTRODUCTION

Nowadays, the digital information over the World Wide Web is growing rapidly. People can access available search engines and portals to enrich themselves about day to day information. People often spend time on searching the specify data rather than reading it due to its huge quantity. Many times, users are unaware of the relevant data due to the time constraint or their current activity does not permit them and hence they remain unfulfilled or uncertain regarding the particular topic. Also, the unrequested information can become distraction from the user's primary goal.

The challenge in the universe of information providers and information users is to call for Copernican revolution shown in figure1 [1] that would place the user in the central position than the information search engine. The Recommender system became an important research area since the mid-1990. Recommender systems are developed to fill the gap between information collection and analysis by filtering all of the available information to present what is most valuable to the user. Information Retrieval, and mainly the sub domain Document Retrieval or Text Retrieval, are concerned with developing technology to find documents that are relevant to a given query.

Document recommendation can be useful in meeting analysis, recommending research papers, online document finding, e-learning system, social tagging to find out online resources which countenance user to get specific information at ease. Figure 2 shows the general structure for document recommendation in meetings. Recommending documents to users mostly in the middle of conversation helps people find relevant information without distracting the conversation. Conversation input contains much number of words than the query. A set of terms from the document called Keywords describes the topic and whole content of the document. Extraction of such keywords is widely used in data mining. Keywords extraction is widely used in summarization of the text, in a tag-based recommendation system as tags. Keywords can also summarize the document collection and they can be used in query expansion and also be used in research paper recommendation. Though a document typically concerns numerous topics in different extents but is relates to a particular topic. Topic modelling examines a set of documents, based on the statistics of the words in each, and decides what the topics might be [2]. The document consist of the number of feature set and most of the feature set like preposition , article don not play any role in representing the document. The feature selection and term weighting based on feature importance in recognizing particular category are very important in deciding result. The weighting scheme is either unsupervised or supervised. Many term weighting scheme are being used like binary, TF, TF-IDF are unsupervised, CH2, PROB, VRF, RF and ICF are supervised [3]. Selecting proper term weighting scheme improves the efficiency, and accuracy of the retrieval result.

The paper is organized as follows. First it defines the state of art: just in time retrieval agent and keyword extraction where various existing just-in-time retrieval agents has been review. It also examines various keyword extraction methods and the term weighting scheme. Table 1 and table 2 describes the comparative analysis of the just in time retrieval agent and term weighting scheme respectively. After the analysis done final conclusion is reported.

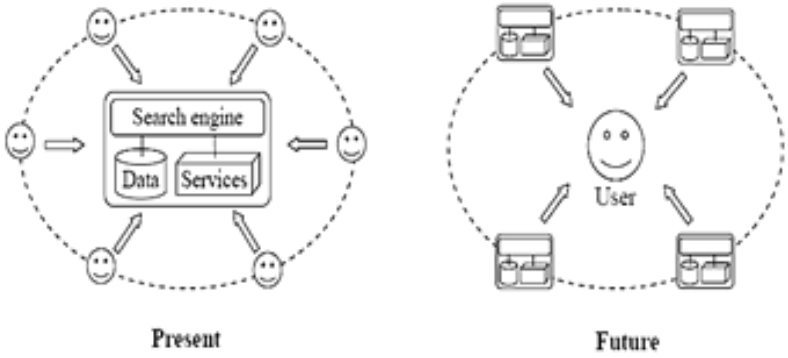

Fig.1Copernican revolution [1]

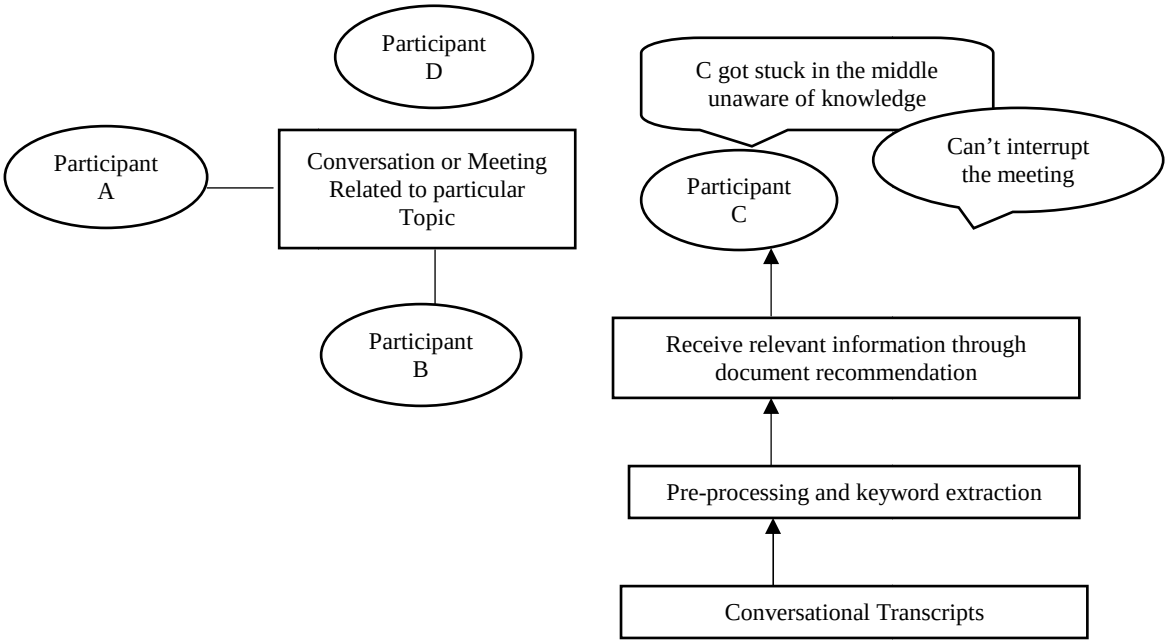

Fig. 2 General Structure for document recommendation in meeting analysis

STATE OF THE ART: JUST IN TIME RETRIEVAL AGENT AND KEYWORD EXTRACTION

Just-In-Time Information Retrieval agents is a class of software agents that proactively present potentially valuable information based on a person's local context in an easily accessible yet non-intrusive manner[4].For example, taking notes during the conference where laptops is not allowed, JITR extract implicit queries from the words and suggest document. Extraction of keywords is related to ranking the words and determining the right weight structure will help in retrieval of relevant result.

A) JUST IN TIME RETRIEVAL AGENT

Query Free Search

The first system for document recommendation was the Fixit system [5] which introduces query-free information retrieval, in which information relevant to user is offered without explicit request. This system integrates with a pre-existing full text database of maintenance manuals. A table of contents or Toc is maintained where each node corresponds to the topic in the documentation. Retrieval is done by use matching query where tree is parsed from higher node to the lower node and returns a particular topic on a single path. It identifies database topic that connects faults and symptoms directly and provides additional support information related to current state.

Remembrance Agent

The Remembrance Agent display the list of relevant document related to user's context and run without user intervention. The RA is broken into two parts. A front end continuously watches what the user types and reads, and

sends this information to the back end. The back end finds old email, notes files, and on-line documents which are somehow relevant to the user's context. This information is then displayed by the front end in a way which doesn't distract from the user's primary task [6]. The desktop RA runs in background and continuously suggests the related documents, old emails, papers at the bottom of screen, related to the context user type in the word processor. The RA is more advantageous in wearable computer. It suggests the document related to the notes taken during conference.

Emac, a UNIX based text editor displays a one line suggestion based on the latest word type and also shows the numeric rating to describe the relevancy of document. For example, one of the suggestion lines regarding the email would be the information about the sender. It makes use of the SMART information retrieval program. Margin notes is a JITIR agent that automatically rewrites Web pages as they are loaded, adding hyperlinks to personal files. As a Web page is loaded, it adds a black margin strip to the right of the document and compares each section of the document to pre indexed e-mail archives, notes files, and other text files, based on keyword co-occurrence. A small suggestion note is included if one of these indexed files is found to be relevant. The RA next step was a wearable assistant called Jimminy brings information based on a person's physical environment, location, time and subject of conversation [4].

The Watson retrieve related document while user is writing or searching over web. It follows some heuristics in constructing an algorithm such as removing stop words, frequently used words, valuing emphasized words, word that come at beginning rather than end and ignoring intentionally deemphasized words and words occurring in the navigation bar of web page [7].

Real Time Agent

Other real time assistants based on conversation is the Ada and Grace are twin virtual guides they interact with user to answer their explicit information needs. Many researchers have developed the automatic content linking device (ACLD) which is a just in time retrieval agent for conversational surrounding specially used in the meeting. It focuses on dialogue act recognition, topic segmentation, subjectivity, and summarization. The keywords are extracted from ASR and summaries describes the content of the meeting [1]. This all motivates towards the diverse keyword extraction from the conversation so that which consists of words than the query so that at least one document would be recommended. Keyword extraction and clustering processed the transcript to find diverse keyword by using LDA. LDA is one of the topics modeling method which help in deciding the main topic. Keyword clustering is also performed to group topics with similarity and ranked by each topic based on their probability [8]. Many studies address term clustering. **Similarity based clustering** group the terms with the same role, e.g. Monday, Tuesday, or happy, fun. **Pairwise clustering** yields relevant terms in the same cluster e.g. Teacher, Prof., HOD, and Principal [9].

B) KEYWORD EXTRACTION

Keywords in the document describe the value of the document. Keyword extraction uses the term weighting scheme to perform the retrieval. Three levels can be used for word assessment.

1) **Corpus Level:** It searches the words that are important, common than expressive but not too common, and not rare. For eg, the words like 'Simple plan', 'Metallica' , 'Green day' are in concern than the words like 'music', 'events', 'bands'. In this the word frequency is used in corpus level.
2) **Cluster Level:** It emphasize words that occur often within a set of similar documents and rarely in other documents. Words having high category probability. For e.g., the band name and member ('Simple Plan', 'Pierrebouvier'), team and member ('Indian', 'SachinTendulkar').Clustering helps in group the similar topics.
3) **Document Level:** To find the keywords in the document. It takes use of the informative worfd found either in corpus level or cluster level, or in both [10]**.**
4) **Related work for keyword extraction**

The first keyword extraction was based on word frequencies which neglects the topicality. Wikify system [11] selects keywords and links them to the external information providers such as online encyclopaedia which gives the detailed information regarding the selected keyword. Research paper recommendation helps researchers keep track of their research field [12] describe the keyword extraction method for the recommendation , in this The keywords are extracted either by keyword extraction algorithm which makes use of the keyword section in the paper or by keyword selection algorithm if the keywords are not present by making the use of title and content. It uses the cosine similarity to calculate degree of similarity.

Some approaches do not use supervised learning but uses some statistics term. Key Graph is based on term co-occurrence, graph segmentation and clustering. It does not make use part-of-speech tags, large corpus, nor supervised learning. It assumes that each cluster hold keyword and find important cluster from document [10], an approach that uses a single document as its corpus is discussed which use the co-occurrences of frequent terms to

evaluate if a candidate keyword is important for a document. The evaluation is done using Chi-squared measure. All of these approaches are designed for longer documents and they rely on term frequencies [13].

To consider the dependencies among selected words, Keyword extraction based on page rank uses Wordnet and graph where Page Rank. Firstly, a text is represented as an undirected weighted semantic based on wordnet where node defines synsets and edges define s relation of nodes and edge is weighted by the relatedness of connected nodes synsets. The second step is to disambiguate words referring to UW-PageRank score, co-reference sense priority and the frequency priority. Then, graph is pruned leaving only the precise synsets nodes. Finally, UW-PageRank is used again to extract key synsets node in the graph, and the corresponding words are assigned as the keywords [14].

Maximum marginal relevance (MMR) is a widely used algorithm for meeting summarization extracts all n grams consisting of content word like adjective nouns from the wordnet dictionary and removing the stopwords and then reweighting all the words. It removes noise from the meeting transcript remove n-grams which appear only once or are fully enclosed by longer n-grams sharing the same frequency [15].

CONCEPTUAL ANALYSIS

As reported in just in time retrieval agent section, Table1 shows the short comparative description among the existing just in time retrieval. Fixit system is a diagnostic system has an advantage of being query free search and offers the users a maintenance manual analyzing the faults and symptoms but was limited only to the some stored database manuals. Three remembrance agent Emac, Margin notes and Jimminy overcome the limited data storage. Emac suggest the document from the last words and no human pre-processing of the documents being indexed is required but there is lot of difference between the relevant information and useful suggestions and it has many design issue. Margin notes works with the web. Emac and Margin notes works with the person computational surroundings while jimminy takes the advantage of the person's physical environment but the disadvantage is that sensor may not give accurate result.

Watson takes the benefit of the emphasized fonts but work with the associated URL, and takes more space and it is relatively unstructured as it uses third party search engine. Many real time conversational retrieval systems such as Ada and Grace listen to user's spoken word and provide information. MindMeld enhance the retrieval by adding user's location. ACLD make use of the ASR engine recognizes the words spoken during meeting and flavors the retrieval by adding additional information like who is the speaker in meeting, who agree or disagree the discussion. Diverse keyword extraction takes the advantage of diverse keyword extraction obtained through ASR and topic modeling. Clustering the similar group of topics give the relevant result, but this system neglects n-gram words.

Table-1 Comparative Study of Existing Just in Retrieval System

Parameter	Remembrance Agent	Working	Data collection	Merits	Demerits
Fixit	Desktop Based	Match text against TOC (Table Of Content)	Full-text database maintained by RICOH.	Query free information offered without explicit request.	Retrieval is limited , from manual stored in database
Jiminy	Wearable Based	Retrieve based on person physical surrounding using Savant.	Annotated corpus of nodes and local context (physical surrounding)	Along with the buffer information it also uses data returned by physical sensors.	Sensor can wear out. Plan for small screen
Watson	Wearable Based	Based on emphasis fonts while writing or browsing	Domain-specific third party search engine.	Current local context rather than user profile or historical information.	1. relatively unstructured 2. related URL. 3. Takes up more space.
Mind meld	Conversational Based	Uses ASR for keyword extraction.	It uses schema.org and Open Graph tag.	Add user's GPS information to the keyword.	Speech recognition is inaccurate because of noisy data.
AMIDA ACLD	Conversational Based	Uses ASR in meetings for document recommendation.	Corpus of the 100 hour meeting recording transcripts.	Detects agreement and disagreement in meetings.	1. Graphical layout of interface is small. 2. Additional functionalities can be suggested.
Keyword extraction and clustering	Conservational Based	Diverse keyword extraction and clustering of ranked keywords.	Domain specific meeting transcripts are used.	Diverse keyword retrieved at least one relevant document.	N grams words are neglected.

Various keyword extraction methods and the term weighting scheme used for the extraction are discussed separately above. Table-2 shows different term weighting scheme. Binary term weighting is simplest in term of time complexity and implementation but many methods has been proposed to improve the accuracy .Term frequency TF-IDF term weighting scheme ranks the keywords and select the highest one. It is good and simple to compute But TF-IDF has some limitations as sometimes; there is no corpus for computing IDF. IDF provides high value to rare terms and low value to common terms.TF is viewed independent and may decrease the precision [16]. TFIDF was unable to uncover the latent semantics associated with the corpus.

To obtain the lexical semantic information various methods have been proposed such as the manually constructed wordnet or from Wikipedia or from the topic modelling technique such as LDA, LSA or PLSA or many supervised machine learning methods or other algorithm such as common random fields , naive Bayes. With LDA, Latent association between documents and multiple topics, and association between topics and keywords are identified [17].CHI^2 do not express the terms positive or negative impact. TF relevant frequency differentiates the term on the basis of positive and negative factor but it works well for the binary classifiers only. Positive impact of the term can be used to calculate its negative effect on other category. PIF term weighting scheme provides high accuracy.

Table-2 Comparative Analysis of Various Term Weighting Schemes

Term	Description	Scheme	Merits	Demerits
Binary	Binary Feature Representation	Unsupervised	Simplest in terms of implementation and time complexity	1.Result contains either too few or too many document 2.no ranking of documents
TF	number of times a term occurs in a document	Unsupervised	Obtain larger scatter of features and accumulates information at fast rate	Common words have higher term frequency which results in Less Recall
TF-IDF	Words in corpus and in single document	Unsupervised	Best known term weighting scheme in information retrieval. Used as a simplest ranking factor.	1. Can't compute IDF if no corpus is available. 2.TF is viewed independent and may decrease the precision 3. Latent semantic among corpus is uncovered.
CHI^2	measure the correlation between feature and class	Supervised	Performs faster than the vector classification method.	From a statistical point of view chi [2] feature selection is problematic Do not express the term's discriminating power.
TF-RF	Differentiate documents in the positive and negative categories	Supervised	1.Robustness 2.performed consistently on data collections with either skewed or uniform category distribution	Only binary classifiers are suitable for this method.
PIF	Positive effect of a feature can be used to measure its negative effect for other categories.	Supervised	Higher accuracy	High time complexity

CONCULSION

With digital data increasing exponentially there is a great need for efficient and fast text classifiers. Term weighting and feature selection are two of the most important steps in building a good text classifier. The task of text classification is overtaken by supervised term weighting schemes for their high accurate results. However they are computationally very costly. Thus the use of a novel weighting scheme Positive Impact factor (PIF) can be an another approach , wherein Positive impact of a feature to a category can be used to calculate its negative impact for other categories.

Here a comparative study is made of different Just in Time Retrieval agent along with diverse keyword extraction methods. AMIDA proposed ACLD is good as it tags the speaker as well as audience speech by using sensor networks. Among the keyword extraction techniques, LDA algorithm is worthy as it performs association between multiple topics and keywords. PIF term weighting scheme can be used for ranking the keyword which will retrieve the relevant document. A complete study suggests that the research area focus on working with the n-gram words.

REFERENCES

[1] A Popescu-Belis, E Boertjes, J Kilgour, P Poller, S Castronovo, T Wilson, A Jaimes and J Carletta, The AMIDA Automatic Content Linking Device: Just-in-Time Document Retrieval in Meetings, *Machine Learning for Multimodal Interaction*, **2008**, 272-283.
[2] *https://en.wikipedia.org/wiki/Topic_mode,* **2015**.

[3] M Emmanuel, Saurabh M Khatri and Ramesh Babu, A Novel Scheme for Term Weighting in Text Categorization: Positive Impact Factor, IEEE International Conference on System, Man and Cybernetics, **2013,** 2292-2297.
[4] B Rhodes and P Maes, Just-In-Time Information Retrieval Agents, *IBM System Journal*, **2000**, 39 (34), 685-704.
[5] P Hart and J Graham, Query-Free Information Retrieval, *IEEE Expert*, 1997, 12 (5), 32-37.
[6] *http://alumni.media.mit.edu/~rhodes/Papers/remembrance.html*, **2015**.
[7] J Budzik and KJ Hammond, User Interactions with Everyday Applications as Context for Just-In-Time Information Access, *Intelligent user interfaces* , **2000**, 44–51.
[8] Maryam Habibi and Andrei Popescu-Belis, Keyword Extraction and Clustering for Document Recommendation in Conversations, *IEEE/ACM Transaction on Audio, Speech and Language Processing*, **2015**, 23 (4), 746-759.
[9] Y Matsuo and M Ishizuka, Keyword Extraction from a Single Document using Word Co-Occurrence Statistical Information, *International Journal on Artificial Intelligence Tools*, **2004**, 13 (1), 157–169.
[10] Mika Timonen, *Term Weighting in Short Documents for Document Categorization, Keyword Extraction and Query Expansion*, Department of Computer Science, University of Helsinki, Finland, **2013**.
[11] A Csomai and R Mihalcea, Linking Educational Materials to Encyclopedic Knowledge, *Artificial Intelligence in Education: Building Technology Rich Learning Contexts That Work*, **2007**, 23,557–559.
[12] Kwanghee Hong, Hocheo Jeon and Changho Jeon, Personalized Research Paper Recommendation System using Keyword Extraction Based on User Profile, *Journal of Convergence Information Technology*, **2013**, 8 (16), 134 – 138.
[13] J Wang, J Liu and C Wang, Keyword Extraction based on Pagerank, *Advances in knowledge discovery and data mining (PAKDD)*, **2007**, 857–864.
[14] K Riedhammer, B Favre and D Hakkani-Tur, A Key Phrase based Approach to Interactive Meeting Summarization, *IEEE Spoken Language Technology Workshop*, **2008**, 153–156.
[15] M Habibi and A Popescu-Belis, Diverse Keyword Extraction from Conversations, *Association for Computational Linguistics*, **2013**, 651–657.
[16] Rohit Nagori and G Aghila, LDA Based Integrated Document Recommendation Model for e-Learning Systems, *IEEE Emerging Trends in Networks and Computer Communications (ETNCC)*, **2011**, 230-233.