An Art of Review on Conceptual based Information Retrieval

P. Mahalakshmi
Research Scholar, Department of Computer Science and Engineering, B.S. Abdur Rahman Crescent Institute of Science & Technology, India.

N. Sabiyath Fathima
Associate Professor, Department of Computer Science and Engineering, B.S. Abdur Rahman Crescent Institute of Science & Technology, India.

Received November 10, 2020; Accepted December 14, 2020
ISSN: 1735-188X
DOI: 10.14704/WEB/V18I1/WEB18026

Abstract

Basically keywords are used to index and retrieve the documents for the user query in a conventional information retrieval systems. When more than one keywords are used for defining the single concept in the documents and in the queries, inaccurate and incomplete results were produced by keyword based retrieval systems. Additionally, manual interventions are required for determining the relationship between the related keywords in terms of semantics to produce the accurate results which have paved the way for semantic search. Various research work has been carried out on concept based information retrieval to tackle the difficulties that are caused by the conventional keyword search and the semantic search systems. This paper aims at elucidating various representation of text that is responsible for retrieving relevant search results, approaches along with the evaluation that are carried out in conceptual information retrieval, the challenges faced by the existing research to expatiate requirements of future research. In addition, the conceptual information that are extracted from the different sources for utilizing the semantic representation by the existing systems have been discussed.

Keywords

Ontology, Information Retrieval, Conceptual based Information Retrieval, Semantic Search.

Introduction

One of the main issue in a normal keyword based search is that searching for documents in the huge amount of information. The aspect that need to be considered for keyword based search is either capturing the user need or the meaning of words or phrases (or the context) confined in the documents.

Most probably the word used in search are of one or few related terms that lead to exponential increase in the resultant web pages. Thus initiating ambiguity that has been occurred due to multiple meanings of query terms leading to retrieval of mismatch web pages. The keyword search systems are less equipped in handling with word-based linguistic occurrences, for instance polysemy or synonymy, and avoid the interpretation of relations that exists between the search terms.

Search engines explores innovative features that would enhance the representation of both the documents and the query. This paved the research gap to redirect the keyword based models, to the notion of search by meaning known to be conceptual search. This can be obtained by the semantic exploration of documents and queries given to the search system.

In this, sentence level semantic has been determined and its associated concept that represents the information acquired by the sentences are used to index the documents and expand the given user queries. The textual representations varied from the normal keyword scenario to the semantic approach thus acquiring the statistical information pertaining towards the retrieval of documents.

The other major challenge is the ambiguity nature of the word that exists as a query term, usually term to be polysemy and synonymy problem produces irrelevant results (Egozi et al., 2011). In addition, insufficient of terminologies and the coarse granularity of concepts are the issues relating to the concept based information retrieval systems.

The polysemy and synonymy problem can be solved by using word disambiguation method (Voorhees 1994, Agirre et al., 2014), latent semantic analysis (LSA) (Deerwester et al., George et al., 2017) and local document analysis (Xu and croft).

The disambiguation of the words in the input query has been determined by expanding the query with similar words, and thus improves the recall of the information retrieval (IR) system. The query has been disambiguated using Machine Readable Dictionaries, semantically disambiguated (tagged) training corpus, contextual information from raw corpora in related to machine learning techniques. The LSA approach extracts and represents the contextual implication of words using statistical computations applied to the corpus.

The local document analysis determines the top ranked documents produced by the query thus providing contextual information for the given query. Latent Semantic Analysis is a

theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text ”Latent Semantic Analysis is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text” Latent Semantic Analysis is a theory and method for extracting and representing the contextual-usage meaning of words by statistical computations applied to a large corpus of text”.

The issues related to insufficient terminologies and the coarse granularity can be tackled by modifying the explicit semantic analysis (ESA), thus obtaining concept that are not available in the knowledge source (Liu et al., 2017). The ESA approach computes the semantic relatedness of the given query with the web repository (Wikipedia) thus identifying the added terminologies to the query.

This paper investigates existing research on various ways in which semantics incorporated with the retrieval aspects. Some of these strategies include – representation of concepts that have been used by the search systems, approaches that are used in conceptual based information retrieval systems using text mining and machine learning techniques to retrieve the search results, evaluation criteria of the existing systems, the semantic information that has been utilized from the various sources for the existing systems.

Moreover this paper emphasis the challenges that have been tackled by the existing systems and the research directions in conceptual based information retrieval have been discussed.

Outline of the Review on Conceptual Information Retrieval

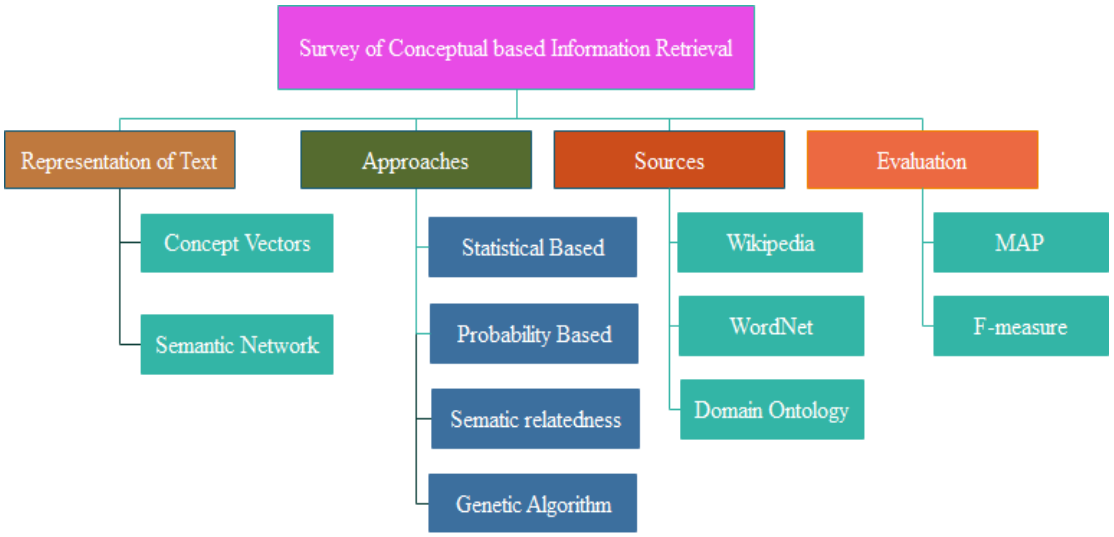

Figure 1 Outline of the Survey on Conceptual Information Retrieval

Representation of Text

Though the search results are obtained by acquiring the concepts that are associated with the query in the documents, text has been considered as different views among the search systems.

Fernandez et. al., [2011] proposed semantically enhanced information retrieval that makes use of documents annotated as concepts. Through SPARQL, query is given by the user that process on semantic index to produces results, based on the semantic entities along with the weighted semantic annotations. Thus enriches the retrieval by bridging the gap between the user query, semantic information about the query and the unstructured textual documents available on the web.

Hsien-Tang Lin et. al., (2012) and Egozi et. al., (2011) used paragraph based features for the search systems. These conceptual retrieval systems make use of documents and the passage set comprising of 50 words to be indexed as concept representation. This emphasizes that the set of concepts present in the passage associates with the query term forms the unique terminologies in retrieving the relevant documents. Light weight ontologies have been considered for representing the documents and the queries as concepts in the retrieval system (Dragoni et al. 2012).This redefines the representation on the recognised machine-readable dictionary (MRD) that moderates the repetition of text commonly confined in a concept-based document representation.

The other view of the text has been considered through semantic network. It is also defined to be associative network that acquires knowledge representation exploring the associated semantic relations between concepts in definite domain of knowledge. This is achieved by edge labeled directed graphs where concepts are represented as vertices and relations are represented as edges. A semantic network is a structural form of predicate logic that acquires information for its representation. Thereby exploiting the relations between the concepts of a domain ontology that has been considered in scoring the relevancy of the document (Gyeong June Hahm et al., 2015). Thus document relevance score has been computed by the domain ontology, the semantics of a document conveyed by a graph (called Document Semantic Network) and, relation-based weighting schemes emphasising the retrieval results.

In another approach search has been implement through the conceptual graphs (SSCGs) (Zhangjie Fu et al., 2017), extracting the predominant and simplified topic sentences from documents and converted to conceptual graphs (CGs). The obtained CGs are mapped CGs

to vectors and ranked the returned results based on "text summarization score". The enhancement in dynamic semantic network and prolonged semantic net based on WordNet has been proposed by Jiang (2020) for semantic information retrieval. The weighted dynamic semantic network (WDSN) has been constructed with the labeled dynamic semantic network (LDSN) for computing the semantic relatedness of the concepts.

Approaches in Concept based Information Retrieval

The concept based information retrieval has evolved using various statistical approaches involving vector space model (Salton 1983) and probabilistic model (Belkin and Croft 1992). The statistical information relates to the term frequency for finding the relevance of document for the given query (Zhangjie Fu et al 2016, 2017), (Gyeong June Hahm et al 2014, 2015), (Francesco Colace et. al., 2015). Latent Semantic Indexing reduces the dimension of index in terms of query and documents used for search systems (Deerwester et al., 1990).

In addition to the concepts, the relation associated with the concepts influences the effect of results produced in the search systems. Fabrizio Lamberti et. al., (2009) explores that the relations among concept in a semantic annotations defined in ranking strategy produces the effective results in semantic web search engines. The underlying assumptions made to the semantic annotations is that "for each concept specified in the query, should have to be characterized by relations with at least another concept". The relation based page rank algorithm estimates the ranking criterion by the probability, that measures between the graph based description of the user query using ontology and the tagged page enclosing queried concepts. The relation based page rank relies on the knowledge of the user query, the pages that are to be ranked and the underlying ontology. Dragoni et al., signifies the representation of terms in the documents computes the appropriate weight to the terms that were indexed and estimates the concepts available in the document and the query. If the concept is not exist in the ontology then the information in the document representation search result would be left out.

In another approach of concept based information retrieval for a specific domain has been proposed by Hsien-Tang Lin et. al. (2012). The system uses Onto Passages for specific search engine. The methodologies of classical IR models have been used to construct their appropriate indices. The methodologies used are of vector space model, probabilistic model, and language model. The limitations of the model is when multi-topic assigned to the particular passage, this approach is inconceivable to produce a paragraph with

multiple topics or concepts. This approach is a time consuming and needs large space for storage.

The usage of thesauri, local document analysis (co-occurrence of terms) and latent semantic analysis that are used to solve the issues related to keywords, by the conceptual information retrieval systems. Egozi et. al., (2011) introduced an Explicit Semantic Analysis (ESA) method for conceptual information retrieval that interprets the features of concepts from information sources. In this additional to conceptual index of a document, each passage has been represented as concept vectors are also indexed and ranked separately and hence differentiates the original document's relevance.The query concepts has been optimized with the feature selection done in the selective-ESA approach. The selective-ESA approach uses the pseudo relevance feedback with selecting the features based on the Information Gain, Intelligent Information Gain and Rocchio Vector of the obtained candidate query concepts. Thus ESA approach enables the selection of features so as to produce more relevant search results. The modified ESA approach (Liu et. al., 2017) uses concept expansion and reranking algorithm to retrieve the results efficiently.

Binbin Yu proposed an ontology based information retrieval consisting of document processing and document retrieval with agent modules. In this, the genetic algorithm has been used for calculating the weighting factor thus enabling the optimum value produce to the concept occurred which intern determines the relevance of the document.Semantic similarity measures has been carried out between the query and the documents for effective retrieval of the results. In another approach of enhancing the information retrieval with different knowledge sources (Jiang, 2020) explores the semantic relatedness between the query and the documents to retrieve the results. Various knowledge sources for example, Wikipedia, Word Net, Description Logic (DL) Ontology have been used to compute the relatedness between the concepts.

Sources of Knowledge Base

Traditionally the concept based information retrieval rely on various sources to acquire information. Those information has been obtained through the following aspects:

Wikipedia – general domain knowledge source. A wide range of topics in Wikipedia is a disadvantage for domain specific retrieval as it retrieves irrelevant topic.

Word Net- semantic lexicon comprises of set of synonyms called syn sets. Word Net provides relations between synsets such as hypernymy, hyponymy, holonymy and meronymy. Ambiguous words produces multiple syn sets.

Ontology – domain specific concepts and relations are represented the hierarchical structure. Relevant documents pertaining to the concepts can be retrieved. The different forms used under ontology format are OWL (Ontology Web Language), RDF (Resource Description Framework), DL (Description Logic) Ontology. BabelNet, YAGO, DBpedia are some other knowledge sources.

Evaluation Aspects of Concept based Information Retrieval

Egozi et. al., (2011) framed out a MORAG system that highlights both Bag-of-Words and ESA method to perform the conceptual based information retrieval. Wikipedia as the knowledge source is used to represent the concept vectors of document for indexing and retrieval. The system has been tested with the datasets of TREC and Robust following the evaluation metric of MAP. Soner Kara et. al., (2012) proposed the ontology-based framework for the extraction and retrieval of semantic information in limited domains. The system consists of a crawler module, an automated information extraction module, an ontology population module, an inference module, and a keyword-based semantic query interface. The system has been tested for a tiny Knowledge base of soccer domain. The enhanced ESA approach (Liu et. al., 2017) has been implemented for Building Information Modelling (BIM) product model retrieval in construction industry and evaluated for the performance metric using MAP.

The arithmetic mean of the average precision values over a set of n query topics (Beitzel et. al., 2009) is defined as the Mean Average Precision (MAP) and is given by

$$MAP = \frac{1}{n}\sum_n AP_n \qquad (1)$$

AP denotes the average precision value for the given topic. MAP is most of the widely used metric for calculating the relevancy of the search results.
F-measure determines the weighted harmonic mean of precision and recall.

$$F = 2\frac{P.R}{P+R} \qquad (2)$$

P denotes the precision conveying the number of documents retrieved. R denotes the recall conveying the number of relevant documents retrieved in the search results.

Jiang (2020) defined macro averaging and micro averaging metric for testing the search results. Macro averaging defined to be the un weighted mean across the queries which is known to be query centric measure. Micro averaging calculated from the sum of per-query contingency tables and is known as document centric measure.

Various Aspects of Conceptual based Information Retrieval

Table 1 Different strategies on Conceptual based Information Retrieval

	Different means of Conceptual based Information Retrieval			
Representation of text	Concepts	Semantic Network	Conceptual Graphs	weighted dynamic semantic network
Approaches	Statistical methods (vector space model and probabilistic model)	Language Model	Explicit Semantic Analysis	Genetic Algorithm
Sources	Wikipedia	WordNet	Domain Ontology	DBpedia
Evaluation	MAP	f-measure	Macro averaging	Micro averaging

Table 1 illustrates the different strategies involving in the conceptual based information retrieval. The text have been used in the form of concepts, conceptual graphs and weighted dynamic semantic network. Considering graph model for a data mining application seems to be time consuming. The approaches used for the information retrieval model are statistical based methods, language model, explicit semantic analysis and genetic algorithm. The challenges faced by the existing conceptual based information model to be rectified by using recent model in machine learning algorithm in retrieving the relevant documents. Due to the increase in the web content, the information regarding to the particular domain to be complete so that in sufficient terminology issue can be overwhelmed. The conceptual based retrieval system with the outcomes have to be evaluated in identifying the relevant documents.

Discussion

The basic flow of a traditional conceptual based information retrieval system is depicted in figure 2. The documents are collected from the web and are processed. The peculiar document processing involves semantic annotation, concept (feature) extraction or contextual graph representation based on the knowledge base. The extracted concepts are indexed using inverted index technique. The document collection, document processing and indexing are carried out as offline process. When the user enters the query for the search, the query is processed resulting in query expansion. The expanded query is given to the index and produces the document. The similarity of query and the returned document results in the appropriate position of the document in the ranked documents.

The query process and finding the similarity between the query and the indexed document are carried out as online process.

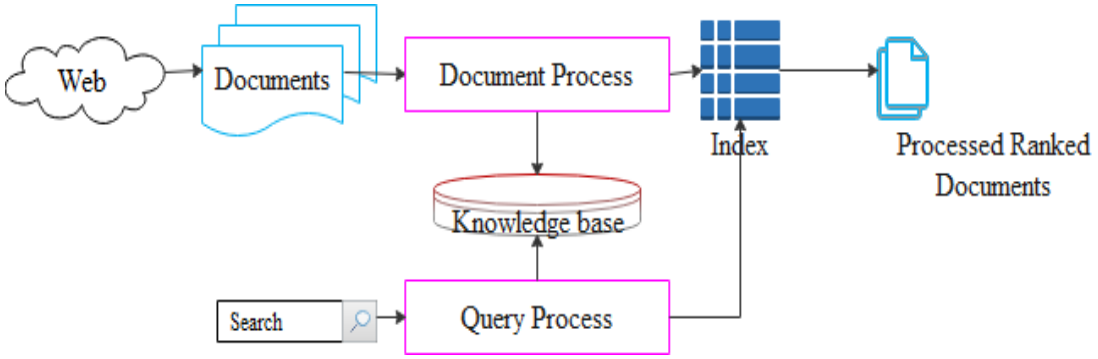

Figure 2 General flow of conceptual based Information retrieval

In concept based information retrieval system, the primary challenge is of collecting the documents from the web as the documents can be of any form. The collected document has to be processed in terms of document processing that acquires the knowledge base information. The knowledge base must able to have the wide range of information related to the concept and also it must be an up-to date resource. To have an up-to date concepts the knowledge base has to be enhanced with the concepts and the associated relation pertaining to the domain. Through the acquired information the documents have to represent either through concept vectors or contextual semantic network that can rely on the retrieval system. The represented concepts have to be indexed providing additional information about the concepts. The index should able to provide a more refined information about the documents. The query given by the user has to be expanded with the additional information acquired from the knowledge base source. The expanded query concept has to be searched in the index that can able to provide the set of documents. The similarity measure or any other approaches have to be applied to yield the relevant documents at the search results. The ranking algorithm should able to determine the similar set of documents to the query and have to locate at the appropriate position in the search results. The whole concept based information retrieval system has to be automated with the advance machine learning techniques to produce the relevant search results.

Conclusion

The survey on concept based information retrieval has been analysed under several aspects of the retrieval system related to the representation of the documents, availability of knowledge sources, approaches that are handled to retrieve the results and the evaluation measures used in the retrieval system. The major criteria behind the conceptual based information retrieval will be of the knowledge sources where in the information

regarding to the particular domain is obtained. The knowledge sources should be of complete and to be up-to date which could be challenge for having domain knowledge. In addition, the issues related to the conceptual based information retrieval system has been discussed. This paper lay out the future direction of the research where the up-to date knowledge base is required and advanced machine learning techniques that have to be adapted for the retrieval system. The process of extracting concepts and relations have to be automated to enhance the knowledge base by adapting machine learning techniques. The validation of the knowledge base have to be done ensuring the appropriate concepts and relations are added to the domain. The constructed knowledge base can be used in data mining applications, emphasizing more on concepts.

References

Jiang, Y. (2020). Semantically-enhanced information retrieval using multiple knowledge sources. *Cluster Computing*, 1-20.

Yu, B. (2019). Research on information retrieval model based on ontology. *EURASIP Journal on Wireless Communications and Networking*, *1*, 1-8.

Furnas, G.W., Deerwester, S., Durnais, S.T., Landauer, T.K., Harshman, R.A., Streeter, L.A., & Lochbaum, K.E. (2017). Information retrieval using a singular value decomposition model of latent semantic structure. *In ACM SIGIR Forum, New York, NY, USA: ACM 51*(2), 90-105.

Liu, H., Liu, Y.S., Pauwels, P., Guo, H., & Gu, M. (2017). Enhanced explicit semantic analysis for product model retrieval in construction industry. *IEEE transactions on industrial informatics*, *13*(6), 3361-3369.

Fu, Z., Huang, F., Ren, K., Weng, J., & Wang, C. (2017). Privacy-preserving smart semantic search based on conceptual graphs over encrypted outsourced data. *IEEE Transactions on Information Forensics and Security*, *12*(8), 1874-1884.

Hua, Y., Jiang, H., & Feng, D. (2015). Real-time semantic search using approximate methodology for large-scale storage systems. *IEEE Transactions on Parallel and Distributed Systems*, *27*(4), 1212-1225.

Fu, Z., Huang, F., Sun, X., Vasilakos, A., & Yang, C.N. (2016). Enabling semantic search based on conceptual graphs over encrypted outsourced data. *IEEE Transactions on Services Computing*, 1-11.

Colace, F., De Santo, M., Greco, L., & Napoletano, P. (2015). Weighted word pairs for query expansion. *Information Processing & Management*, *51*(1), 179-193.

Hahm, G.J., Lee, J.H., & Suh, H.W. (2015). Semantic relation based personalized ranking approach for engineering document retrieval. *Advanced Engineering Informatics*, *29*(3), 366-379.

Hahm, G.J., Yi, M.Y., Lee, J.H., & Suh, H.W. (2014). A personalized query expansion approach for engineering document retrieval. *Advanced Engineering Informatics*, *28*(4), 344-359.

Agirre, E., De Lacalle, O.L., & Soroa, A. (2014). Random walks for knowledge-based word sense disambiguation. *Computational Linguistics*, *40*(1), 57–84.

Lu, Y., He, H., Zhao, H., Meng, W., & Yu, C. (2011). Annotating search results from web databases. *IEEE Transactions on Knowledge and Data Engineering*, *25*(3), 514-527.

Kara, S., Alan, Ö., Sabuncu, O., Akpınar, S., Cicekli, N.K., & Alpaslan, F.N. (2012). An ontology-based retrieval system using semantic indexing. *Information Systems*, *37*(4), 294-305.

Lin, H.T., Chi, N.W., & Hsieh, S.H. (2012). A concept-based information retrieval approach for engineering domain-specific technical documents. *Advanced Engineering Informatics*, *26*(2), 349-360.

Dragoni, M., Da Costa Pereira, C., & Tettamanzi, A.G. (2012). A conceptual representation of documents and queries for information retrieval systems by using light ontologies. *Expert Systems with applications*, *39*(12), 10376-10388.

Egozi, O., Markovitch, S., & Gabrilovich, E. (2011). Concept-based information retrieval using explicit semantic analysis. *ACM Transactions on Information Systems (TOIS)*, *29*(2), 1-34.

Liu, L., & Özsu, M.T. (Eds.). (2009). *Encyclopedia of database systems*. New York, NY, USA: Springer, *6*.

Lamberti, F., Sanna, A., & Demartini, C. (2008). A relation-based page rank algorithm for semantic web search engines. *IEEE Transactions on Knowledge and Data Engineering*, *21*(1), 123-136.

Kao, B., Lee, J., Ng, C.Y., & Cheung, D. (2000). Anchor point indexing in Web document retrieval. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *30*(3), 364-373.

Jinxi, X., & Bruce Croft, W. (1996). Query Expansion Using Local and Global Document Analysis. *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, 4-11.

Voorhees, E.M. (1994). *Query expansion using lexical-semantic relations.* In SIGIR'94, Springer, London, 61-69.

Voorhees, E.M. (1993). Using WordNet to disambiguate word senses for text retrieval. *In Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, 171-180.

Belkin, N.J., & Croft, W.B. (1992). Information filtering and information retrieval: Two sides of the same coin?. *Communications of the ACM*, *35*(12), 29-38.

Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407.

McGill, M.J. (1983). *Introduction to Modern Information Retrieval McGraw-Hill.* New York.