Towards Semantically Enhanced Information Retrieval

A.Tulika Narang
Centre of Computer Education, University of Allahabad
Allahabad, India

B.Prof. R.R. Tewari
Centre of Computer Education, University of Allahabad
Allahabad, India

Abstract - The paper discusses information retrieval on the World Wide Web. The focus is to retrieve and find the most useful and relevant. The challenge is to overcome the limitations of traditional keyword based search. A semantic search overcomes the drawbacks of overload and mismatch associated with keyword based search. A semantic search uses ontology and clustering algorithm to perform relevant and valuable retrieval of web documents. The approach is an attempt to improve the information retrieval process. It aims at finding the most useful information from huge amount of data available on the World Wide Web.

I. INTRODUCTION

World Wide Web is a repository of large volume of data. The process of discovering effective knowledge and right information using keyword based methods has various limitations associated. Lack of personalization as well as inability to easily separate commercial from non-commercial searches is among other limitations of today's web search technologies.

The activity of obtaining information resources relevant to information from collection of information resources is termed as Information retrieval (IR). IR can be based on metadata or on full-text indexing. Automated information retrieval systems are used to reduce what has been called "information overload". It deals with the representation, storage, organization of, and access to information items. The representation and organization of the information items should provide the user with easy access to the information in which he is interested. The key objective of an IR system is to retrieve information which might be useful or relevant to the user. The emphasis is on the retrieval of most useful and relevant information [9].

An information retrieval process begins when a user enters a query into the system. Queries are formal statements of information needs, for example search strings in web search engines. In information retrieval a query does not uniquely identify a single object in the collection. Instead, several objects may match the query, perhaps with different degrees of relevancy [11,12].

An object is an entity that is represented by information in a database. User queries are matched against the database information. Often the documents themselves are not kept or stored directly in the IR system, but are instead represented in the system by document surrogates or metadata.

Most IR systems compute a numeric score on how well each object in the database matches the query, and rank the objects according to this value. The top ranking objects are then shown to the user. The process may then be iterated if the user wishes to refine the query [9].

Relevance denotes how well a retrieved document or set of documents meets the information need of the user. It defines to what extent the topic of a result matches the topic of the query or information need. A semantic search based on the representation of content and associated meanings gives a better result as compared to keyword based search. It reduces the distance between the logic representation of the IR

systems and the real one in the user's mind with regards to the formulation of queries. It includes the introduction of ontologies as conceptual framework.

II. BACKGROUND AND RELATED WORK

The advent of computers made possible to store large amounts of information. The process of retrieving useful information from huge collections of data is a major concern. The field of Information Retrieval (IR) was born out of this necessity. The field has grown considerably from simple keyword based search to semantically enhanced ontology driven information retrieval systems.
In its most common form a set of keywords (or index terms) resolve the users' query.
The simple keyword-based query interface has to a great extent contributed to the wide acceptance of the Internet and its proliferation of user-contributed contents. The method identified the connection of data items containing keywords.
But the keywords entered by users may imply different information needs of different users. The limitation holds prominent in cases such as:
Polysemy: "jaguar" as animal vs. "jaguar" as car
Synonymy: "movies" vs. "films".
This causes ambiguity during query processing and leads to unwanted results.The search quality are not as per users' expectations.

2.1 Vector Space model

Vector space model or term vector model is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers, such as, for example, index terms. It is used in information filtering, information retrieval, indexing and relevancy rankings.
In the vector space model text is represented by a vector of terms. The definition of a term is not inherent in the model, but terms are typically words and phrases. If words are chosen as terms, then every word in the vocabulary becomes an independent dimension in a very high dimensional vector space. Any text can then be represented by a vector in this high dimensional space. If a term belongs to a text, it gets a non-zero value in the text-vector along the dimension corresponding to the term. Since any text contains a limited set of terms (the vocabulary can be millions of terms), most text vectors are very sparse. Most vector based systems operate in the positive quadrant of the vector space, i.e., no term is assigned a negative value.
To assign a numeric score to a document for a query, the model measures the similarity between the query vector (since query is also just text and can be converted into a vector) and the document vector. The similarity between two vectors is once again not inherent in the model. Typically, the angle between two vectors is used as a measure of divergence between the vectors, and cosine of the angle is used as the numeric similarity cosine has the nice property that it is 1.0 for identical vectors and 0.0 for orthogonal vectors).

III. SEMANTICALLY ENHANCED DOCUMENT CLUSTERING

A semantic search is the application of natural language processing to support information retrieval, analytics, and data-integration [6].
A semantically enhanced information retrieval overcomes the limitations of keyword based search. A keyword based information retrieval method has limited capabilities to grasp and exploit the conceptualizations involved in user needs and content meanings. On the other hand, a semantic search focuses on meanings rather than literal strings. The focus is on semantics leading to better and accurate results. A semantic method uses ontology to overcome the limitations of keyword-based information retrieval [8].

A semantically enhanced information retrieval comprises of the following essential phases:

- Preprocessing
- Document clustering

3.1 Preprocessing

Preprocessing includes semantic indexing or annotation of Web documents. Semantic annotation is the process of inserting tags in a document to assign semantics to text fragments allowing creating the documents to be processed not only by humans but also automated agents. Annotations are comments, notes, explanations, or other types of external remarks that can be attached to a Web document or to a selected part of a document. Annotations are metadata giving additional information about an existing piece of data. An annotation has properties such as physical location, scope, type.

Meaningful use of any data requires knowledge about its organization and content. Contextual information that establishes relationships between the data and the real world aspects it applies to is called metadata. In other words, metadata is data that describes information about a piece of data, thereby creating a context in terms of the content and functionality of that data. Domain conceptualizations, ontologies or world models provide agreed upon and unambiguous models for capturing data and metadata to which applications, data providers and consumers can refer. Broadly speaking, there are two kinds of metadata - structural and syntactic metadata [6].

Structural metadata provides information about the organization and structure of some data, e.g. format of the document. Semantic metadata on the other hand, provides information 'about' the data for example the meaning or what the data is about and the available semantic relationships from a domain model in which the data is defined.

Ontology is commonly defined as an explicit, formal specification of a shared conceptualization of a domain of interest. Ontology describes some application-relevant part of the world in a machine-understandable way. The concepts and concept definitions that are part of the ontology have been agreed upon by a community of people who have an interest in the corresponding ontology. The core “ingredients” of an ontology are its set of concepts, its set of properties, and the relationships between the elements of these two sets [1, 2].

Ontological structures give additional value to semantic annotations. They allow for additional possibilities on the resulting semantic annotations, such as inferencing or conceptual navigation that we have mentioned before. Ontology gives guidelines for what and how items residing in the documents may be annotated. But ontology based semantic annotation yields in comparison to “free text metadata generation”, the extended set of capabilities also entails some new problems that need to be solved. In particular, semantic interlinkage between document items incurs the difficulty to adequately manage these interlinkages. Essentially, this means that an ontology-based annotation tool must address the issue of object identity and its management across many documents. Also, ontologies may have elaborate definitions of concepts. When their meaning changes, when old concepts need to be erased, or when new concepts come up, the ontology changes. Because updating previous annotations is generally too expensive, one must deal with change management of ontologies in relation to their corresponding annotations. Redundant annotation should be prevented which stem from duplicate pages on the web or annotation work done by fellow annotators.

3.2 Clustering

The clustering process is applied on annotated documents for semantically enhance information retrieval on the web [7]. It is the process of organizing objects into groups whose members are similar in some way. A cluster is a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. It is an unsupervised learning technique [4, 5]. It is one of the essential tasks of data mining. Data mining is a nontrivial extraction of previously unknown, potentially useful and reliable patterns from a set of data. It is the process of analyzing data from different perspectives and summarizing it into useful information [3]. It is a process of knowledge discovery and useful for exploring data. . It is the division of data into groups of similar objects.

IV. CONCLUSION

Semantic search improves search accuracy by understanding searcher intent and the contextual meaning of terms as they appear in the searchable data space. Semantic search systems consider various points

including context of search, location, intent, variation of words, synonyms, generalized and specialized queries, concept matching and natural language queries to provide relevant search results.

REFERENCES

[1] Xiaohui tao, Yuefeng Li and Ning Zhong, "A Personalized Ontology Model for Web Information Gathering", IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 4, April 2011

[2] Hector Oscar Nigro, Sandra Gonzalez Cisaro, and Daniel Hugo Xodo, Data Mining with Ontologies-Implementations, Findings, and Frameworks, IGI Global, ISBN 978-1-59904-618-1

[3] Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques,ISBN 81-8147-049-4

[4] Margaret H. Dunham and S. Sridhar, Data Mining–Introductory and Advanced Topics,Second Impression, 2007, ISBN 81-7758-785-4

[5] Richard J. Roiger and Michael W. Geatz, Data Mining-A Tutotrial-based Primer,First Indian Reprint, 2005, ISBN 81-297-1089-7.

[6] Wanlong LI, Dayou Liu, Shanhong Zheng, Suyun Jiao," A Novel Computational Approach to Concept Semantic Similarity", International Conference on Computer, Mechatronics, Control and Electronic Engineering 2010.

[7] Ernesto William De Luca, Andreas Nürnberger, "Using Clustering Methods to Improve Ontology-Based Query Term Disambiguation" International Journal Of Intelligent Systems, Vol. 21, 2006

[8] Yuefeng Li and Ning Zhong," Capturing Evolving patterns for Ontology based Web mining", International Conference on Web Intelligence, 2004.

[9] Y. Li and N.Zhong,"Web Mining Model and Its Application on Information gathering," Knowledge Based Systems, vol 7, 2004

[10] R. Baeza Yates and B. Ribeiro Neto, Modern Information retrieval, Addison Wesley, 1999

[11] Christopher D. Manning Prabhakar Raghavan and Hinrich Schiitze, "Introduction to Information retrieval", 2008.

[12] Hotho, Maedche, Staab,"Ontology based text document clustering", Kunstliche Intelligenz, 16(4),pp 48-54.