Web People Search Using Ontology Based Decision Tree

Mrunal Patil[1], Sonam Khomane[2], Varsha Saykar[3] and Kavita Moholkar[4]

[1]Department of Computer Engineering, Rajarshi Shahu College of Engineering, Pune-411033, India.
mrunalpatil21190@gmail.com

[2]Department of Computer Engineering, Rajarshi Shahu College of Engineering, Pune-411033, India.
sonamkhomane@gmail.com

[3]Department of Computer Engineering, Rajarshi Shahu College of Engineering, Pune-411033, India.
varsha3636@gmail.com

[4] Lecturer, Department of Computer Engineering, Rajarshi Shahu College of Engineering, Pune-411033, India.
kavita.moholkar@gmail.com

Abstract:

Nowadays, searching for people on web is the most common activity done by most of the users. When we give a query for person search, it returns a set of web pages related to distinct person of given name. For such type of search the job of finding the web page of interest is left on the user. In this paper, we develop a technique for web people search which clusters the web pages based on semantic information and maps them using ontology based decision tree making the user to access the information in more easy way. This technique uses the concept of ontology thus reducing the number of inconsistencies. The result proves that ontology based decision tree and clustering helps in increasing the efficiency of the overall search.

Keywords

Web people search, Ontology based decision tree, Clustering, Semantic information, Efficiency.

1.Introduction

Development of web and its influence on the people and next generation is increasing rapidly. The most common activity done on the internet involves the web people search. According to Samuel Johnson "Knowledge is of two kinds: we know a subject ourselves, or we know where we can find information about it" [6]. For years, these words remained true to those in search for information. Although not always readily available, information about person could be easily found by individuals in directories, indexed by human experts. With the advent of the Internet, however, the situation changed dramatically. Enormous amounts of data are now freely accessible to over 600 millions of users on-line. The question of how to find information of interested person on the Internet is raised by the Web People Search Problem.

Web search is difficult because it is hard for users to construct queries that are both sufficiently descriptive and sufficiently discriminating to find just the web pages that are relevant to the user's search goal. Queries are often ambiguous. To solve the problem various algorithms are been adopted by the current search engines like the google, yahoo etc. Queries are fired to such search engine. The result returned is in the form of rank list of documents along with partial content leaving the job of finding the specific document onto the user. Thus to find the relevant document user has to shift through a large set of irrelevant document.

Our approach exploits the task of searching the people in more precise manner. Web people search clusters the web pages based on the query fired. Clustering is the process of gathering similar objects in one cluster which are different to the objects in another cluster. Web pages having semantic information are grouped in a single cluster. Later, these clusters are presented to the user in form of ontology. Ontology is a set of concepts such as things and relations that are specified in some way in order to exchange information. Firstly input is given to the search engine in form of query. The search engine returns the top k relevant pages as soon as the query is fired. Web pages are retrieved are processed. The pre processed web pages are clusters and then presented using decision tree. The system can be made more generalized for the search of people with the help of filters.

The main objective of this paper is to understand the basic concepts of ontology with particular emphasis on its application to people search problem. For this purpose, the paper contains the review of past work in part II and explains the general concepts behind ontology in Part III, followed by a general description of the system in Part IV. Part V provides the experimental results. Part VI concludes along with the future work and references.

2.Review of the Past Work

Initially, for searching the Page Rank algorithms were used for ranking the search query results. The page rank algorithm was described by Lawrence Page and Sergey Brin. Working of the Page Rank algorithm depends upon link structure of the web pages. The Page Rank algorithm is based on the concepts that if a page contains important links towards it then the links of this page towards the other page are also to be considered as important pages. The Page Rank considers the back link in deciding the rank score. But the limitation is that if new page is inserted between two pages then the crawler should perform a large calculation to calculate the distance vector which is a time consuming process and decreases the performance. Taher Haveliwala in 2002 proposed a Topic Sensitive Page Rank as compared to the original Page Rank for improving the search-query results where a single Page Rank vector is computed using the link structure of the web to compute the relative importance of the page [4].

Hearst and Pedersen showed that relevant documents tend to be more similar to each other, thus the clustering of similar search results helps users find relevant results. In addition, Vivisimo is a real demonstration of this technique. Vivisimo was founded in 2000 by three Carnegie Mellon University scientists who decided to tackle the problem of information overload in web search. Rather than focusing just on search engine result ranking, we realized that grouping results into topics, or "clouds," made for better search and discovery. As search became a necessity for web users, Vivisimo developed a service robust enough to handle the variety of information the everyday web user was after. The result was Clusty: an innovative way to get more out of every search. Clusty was acquired by Yippy, Inc. in May 2010. Yippy queries several top search engines, combines the results, and generates an ordered list based on comparative ranking.

Jargon Went and Jian-Yun Nie proposed a new approach to query clustering using user logs[3]. The principles are as follows. 1) If users clicked on the same documents for different queries, then the queries are similar. 2) If a set of documents is often selected for a set of queries, then

the terms in these documents are related to the terms of the queries to some extent. These principles are used in combination with the traditional approaches based on query contents.

Our proposed local-cluster algorithm considers linkage structure and content generation of cluster structures to produce a ranking of the underlying clusters with respect to a user's given search query and preference. The rank of each document is then obtained through the relation of the given document with respect to its relevant clusters and the respective preference of these clusters.

3.Ontology

Ontology specifies linked concepts and terms and relations among these terms and concepts. Concepts are nothing but the entities which are language independent. Ontology shows how each concept is related what properties it has. The main purpose of ontology based decision tree is to give a more meaningful, descriptive and a readable view of concepts.

Mathematically ontology can be defined Yang *et al*., 2008 [5] as follows:

"An ontology can be defined as an Vector O: = (C, V, P, H, ROOT), where C is the set of concepts, V contains a set of terms and is called the vocabulary, P is the set of properties for each concept, H is the hierarchy and ROOT is the topmost concept. Concepts are taxonomically related by the directed, acyclic, transitive, reflexive relation H belongs to C * C. H(c1, c2) shows that c1 is a subclass of c2 and for all c belongs to C it holds that H(c, ROOT)."

Our goal is to utilize the user context to the search results by re-ranking the results returned from the given query of search engine. The representation of the tree depends upon the user's information access behavior. Semantic information is fundamental part of user context. The best example of ontological approach has proven to be successful in the recommender system which does not consider the domain knowledge.

Ontology consists of hierarchy of classes and sub-classes for object-entity [1]. The clusters will be taken as input which is formed using the lingo algorithm, later arranged in hierarchy.

This could be achieved as follows:

- The topmost concept is the root having intermediate and leaf concepts. If a user wishes for a leaf concept then each cluster starting from the topmost concept will be traversed.
- The probability for each cluster will be calculated and depending on that clusters of user's interest will be accessed. Long with this a threshold value will be maintained.
- If the user heating ratio for a given cluster is greater and also the probability for each cluster is greater than the threshold value then parent is renamed by child name.

Mathematically it can be defined as follows:

Pb= (100*HR)/Total no.of clusters viewed

Where, Pb=Probability

HR=hitting ratio

An example of basic ontology is shown below in the Figure: 1.

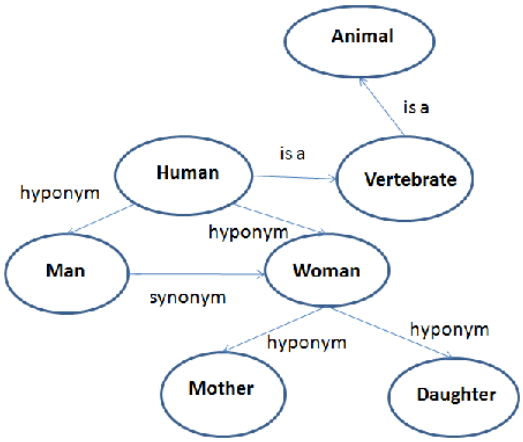

Figure 1: Example of Ontology

Ontology explains a hierarchy of classes, sub classes and their relationships. The above example describes the "HIERARCHY" for a human class. It shows that a human is a vertebrate where the man and woman are synonym and they are hyponym for class human.

Proposed System

The system mainly works in different phases as:

4.1 Search result fetching:

The user submits the query to the search engine. The filters are used to find whether the given query is related to person search or not. Then we get the WebPages of search result lists returned by a Goggle web search engine. So the first search is the conventional met search based on these keywords. These WebPages are analyzed by an HTML parser and the result items are extracted.

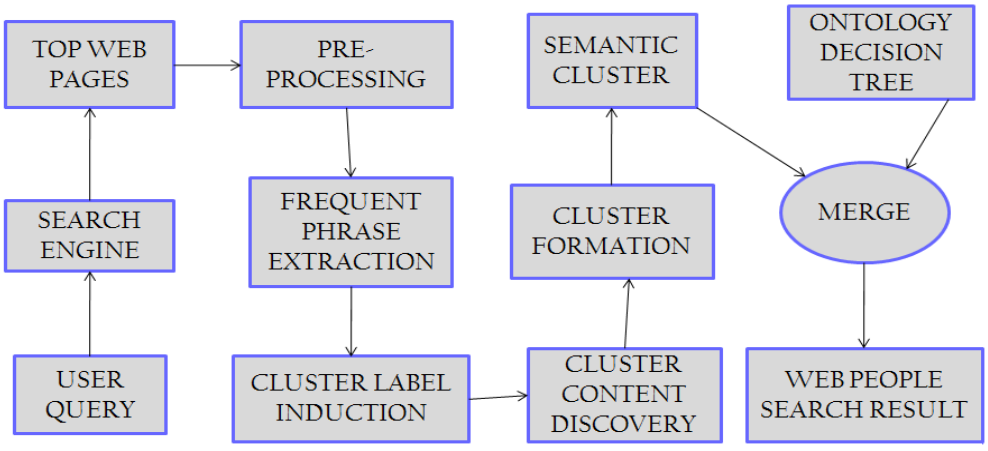

Figure 2: Proposed system

Generally, there are only titles and query-dependent snippets available in each result item. We assume these contents are informative enough because most search engines are well designed to

facilitate users' relevance judgment only by the title and snippet, thus it is able to present themost relevant contents for a given query. Each extracted phrase is in fact the name of a candidate cluster, which corresponds to a set of documents that contain the phrase.

4.2 Ontology based Web People Search:

When designing a Cluster Based Web Search, special attention must be paid to ensuring that both content and description (labels) of the resulting groups are meaningful to humans. There are various algorithms such as K means, K-medoid, Suffix tree clustering, Hierarchical clustering. There are various drawbacks of these algorithms and to overcome these we use the Lingo algorithm. Lingo reverses the process—first attempt to ensure that we can create a human-perceivable cluster label and only then assign documents to it. Specifically, extract frequent phrases from the input documents, hoping they are the most informative source of human-readable topic descriptions. Finally, match group descriptions with the extracted topics and assign relevant documents to them.

Our novel algorithm, Lingo clusters the search results. Unlike other algorithms lingo first discovers the name of the clusters and then evaluate each cluster with appropriate name. Basically lingo consists of five phases. The first phase deals with preprocessing to data. It detects the word boundaries and applies the stemming and stop word removal. The second phase deals with the frequent phrase extraction here a term appearing certain number of times is discovered and combined in all set of documents. Phase three is the cluster label induction. SVD is used to extract orthogonal vectors of the term document matrix, believed to represent distinct topic in the input data [6]. The fourth phase is the cluster content discovery phase. In this phase a Vector Space Model is applied to put the input documents under the cluster label discovered in the previous phase. Highest scoring documents for each cluster are assigned as that cluster's content [7]. Last phase is the i.e. the fifth phase is final cluster formation. Later, ontology based decision tree is referred for rearranging the clusters formed. The steps that will be followed are:

a. The ontology tree is formed according to the hitting ratio. Each object forms here the separate cluster.
b. After the cluster formation, weight is given to each node of the tree and threshold value is maintained.
c. Rearrange the clusters considering the weight given to each cluster and then display them to the users. Procedure that calculates the similarity between objects and clusters that estimate the similarity between clusters and ontology objects are used for this purpose.
d. If desired number of clusters are obtained, then stop else goto step a.

Experiment Results

This section deals with experimental results when applied on web people search using ontology based decision tree. We have searched for the different persons with their relations like Abdul Kalam as the president of India, and listed the results obtained from various search engines The results obtained from our search engine where compared with other search engine likes Yippy, Wink People search engine, etc. Then these results were used to find the overall efficiency and accuracy of the search engines to give relevant results. We found that the accuracy for our search engine i.e. Web People Search Engine is good than other search engines.

The results are listed in the following table:

Table 1: Results Obtained from different Search Engines

Search Engine	**Yippy search engine**		**Wink Search Engine**	**Web People Search**	
Query :	Clusters Obtained	Relevant Clusters	Results Obtained	Clusters Obtained	Relevant Clusters
A.P. J. Abdul Kalam as president	A.P.J. Abdul Kalam (46) Dr APJ Abdul Kalam (12) Speech, APJ Abdul Kalam (5) Photos (7) Sri Lanka (5) Science, Complements Modern (5) Videos (5) Award, Annual (5) Programme (3)Former Indian president APJ Abdul Kalam (2)	A.P.J. Abdul Kalam (46) Dr APJ Abdul Kalam (12) Speech, APJ Abdul Kalam (5) Former Indian president APJ Abdul Kalam (2)	**abdul kalam** (Hyderabad, ANDHRA PRADESH) **Abdul Kalam abdul kalam** (Nellur, ANDHRA PRADESH) • Web & Graphic Designer **Abdul Kalam** (Mumbai, MAHARASHTRA) **ABDUL KALAM** Infosys • Project Manager **abdul kalam** (Bengaluru, KARNATAKA) **aneesha abdul kalam** (KERALA) **Shaik Abdul Kalam**	A.P.J. Abdul Kalam (15) Abdul Kalam Quotes (13) President APJ Abdul Kalam (10) Avul Pakir Jainulabdeen Abdul Kalam Tamil (9) Tamil Nadu (9) A. P. J. Abdul Kalam (7) Latest News (5) Abdul Kalam's Death (3) Security (3) Abdul Kalam Horoscope (2) Madras Institute (2) Notable Scientist (2) United Kingdom LinkedIn (2) Other Topics (34)	A.P.J. Abdul Kalam (15) Abdul Kalam Quotes (13) President APJ Abdul Kalam (10) Notable Scientist (2) A. P. J. Abdul Kalam (7) Latest News (5) Other Topics (34)
Narayana Murthy as CEO of Infosys	Narayana Murthy (52) N. R. Narayan Murthy (7) Famous (6) Phaneesh Murthy (5) Magazine (3) Mysore (3) Chief Mentor. Infosys (3) Blog, Time (3) Committee Report (2) Humbleness At Its Best . Salute … Sunil On Narayan Murthy (2)	Narayana Murthy (5 2) N. R. Narayan Murthy (7) Famous (6) Chief Mentor. Infosys (3) .	Narayana Murthy ***(Mumbai, MAHARASHTRA*• SAP Basis Consultant narayana murthy *(ANDHRA PRADESH)* Narayana Murthy *(Bengaluru, KARNATAKA)*Narayana murthy *(Mysore, KARNATAKA)*Narayana Murthy** Indian Institute of Technology, Kanpur • Mavenir Systems • Development Manager	New CEO (9) Kris Gopalakrishnan (5) Software Professionals (5) Emeritus (4) Asia Society (2) CIOL News Reports (2) Economic Times Corpo-rate Dossier List (2) Infosys with an Equity Capital of Rs.10 (2) NiTiN Jadhav (2) Other Topics (30)	
Sachin Tendulkar as Cricketer	Retired, ODI (6) Sachin Ramesh Tendulkar (8) Photos (8) Test (6) Cricket player (5) God (6) Day cricket (5) Batting,	Sachin Ramesh Tendulkar (8) Test (6) Cricket player (5) God (6) Day	Sachin Ramesh Tendulkar sachin tendulkar My Places: **http://www.mylife.com/displayProfile.do?uid=561065151** sachin tendulkar My	Cricketer in the World (15) Sachin Ramesh Tendulkar is an Indian Cricketer (15) Batsman Sachin (13) Sachin Ramesh Tendulkar Born (11) History of Cricket	Cricketer in the World (15) Sachin Ramesh Tendulkar is an Indian Cricketer (15)

	Centuries (5) Dhoni (3)	cricket (5) Batting, Centuries (5)	Places: **Facebook.com: Sachin Tendulkar** Sachin Tendulkar (Belgaum, KARNATAKA) Indian Agricultural Research Institute • Mineral Foundation of Goa • Program Manager	(10) World Cup (7) Year (7) Sachin Tendulkar Biography (6) Sachin Tendulkar Profile (6) Latest News (4) Sachin Tendulkar in 2011 Cricket World Cup (3) Sydney (3) Australian Batsman Michael Hussey Feels (2) BBC News (2) Cricinfo (2) Other Topics (46)	Batsman Sachin (13) Sachin Ramesh Tendulkar Born (11) History of Cricket (10) Sachin Tendulkar Biography (6) Sachin Tendulkar Profile (6) Latest News (4) Sachin Tendulkar in 2011 Cricket World Cup (3)
Shahrukh Khan as Film Star	Photos (26) Best Actor (10) Bollywood Actor (9) Indian film actor (7) Bollywood Actor Shahrukh Khan (9) Bollywood superstar Shahrukh Khan (7) Global (7) My Name Is Khan (6) Bollywood Film (7) Aamir Khan (6)	Best Actor (10) Bollywood Actor (9) Indian film actor (7) Bollywood Actor Shahrukh Khan (9)	Shahrukh Khan Khan (New Delhi, DELHI shahrukh khan My Places: http://www.mylife.com/displayProfile.do?uid=450737068 Shahrukh Khan **Array** Shahrukh Khan (Bīkāner, RAJASTHAN) Shahrukh Khan (Mumbai, MAHARASHTRA) shahrukh khan (Mumbai, MAHARASHTRA) Karnatak University	Shah Rukh Khan (13) King Khan (9) Born 2 November 1965 (8) International (6) Shahrukh Khan Picture (6) BBC News (5) Embroiled in an Ugly Row on Monday (5) Don 2 SRK (3) Videos (3) Dubai for New Year by Nazia Khan (2) Kareena Kapoor (2) Lady Gaga (2) Leonardo DiCaprio (2) Other Topics (37)	Shah Rukh Khan (13) King Khan (9) Videos (3) International (6) Shahrukh Khan Picture (6) Other Topics (37)

Based on the above results obtained we can find the overall accuracy of the various search engines which is shown by following graph.

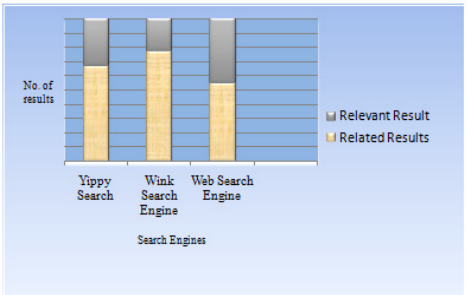

Figure 3: Result Graph

5.Conclusion

Our method exploits a web mining tool, knowledge driven cluster based search system that helps the user to find web information based on individual preferences. Our system also shows that, while it is possible to improve the efficiency of search through ontology method discussed above, it infact works best when operated in conjunction with one another and provide better search result.

References

[1] 'Impact of Ontology based Approach on Document Clustering', *International Applications .Journal of Computer (0975 – 8887)Volume 22– No.2, May 2011.*

[2] 'Web People Search via Connection Analysis' by Dmitri V. Kalashnicov, Zhaoqi (Stella) Chen, Sharad Mehrotra, Member, IEEE, and Rabia Nuray-Turan. *IEEE TRANCTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 20, NO. 11, NOVEMBER 2008.*

[3] 'Lingo: Search Results Clustering Algorithm Based on Singular Value Decomposition' *By Stanis law Osi´nski, Jerzy Stefanowski, and Dawid Weiss, 2004.*

[4] 'Topic Sensitive PageRank' , By Haveliwala.

[5] 'Research on Ontology-Based Text Clustering', By Yang, X., Guo, D., Cao, X. and Zhou, J. (2008) *Proceedings of the 2008 Third International Workshop on Semantic Media Adaptation and Personalization, EEE Computer Society Washington, DC, USA.*

[6] Michael W. Berry, Susan T. Dumais, and Gavin W. O'Brien. Using linear algebra for intelligent information retrieval. *Technical Report UT-CS-94-270, 1994.*

[7] 'Conceptual Clustering Using Lingo Algorithm Evaluation on Open Directory Project Data', By Stanis law Osi´nski and Dawid Weiss , *Institute of Computing Science, Poland University of Technology,2003.*

[8] 'Lingo: Search results clustering algorithm based on Singular Value Decomposition' By Stanis law Osi´nski, Jerzy Stefanowski, and Dawid Weiss. Submitted to *Intelligent Information Systems Conference 2004, Zakopane, Poland, 2003.*

[9] ' Disambiguating People in Search. Stanford Univ*', By R. Guha and A. Garg, 2004.*

[10] 'An algorithm for clustering of web search result'*, By Stanisław Osinski,2003.*

[11] 'Ontology-Driven Induction Trees at Multiple Levels of Abatraction' , By Jun Zhang, Adrian Selvescu and Vasant Honavar. *Artificial Intelligence Research Laboratory Department of Computer Science, Iowa State University Ames, Iowa-50011-1040 USA.*

[12] 'An Empirical Evaluation on Semantic Search Performance of Keyword- Based and Semantic Search Engines: Google, Yahoo, Msn and Hakia Duygu Tümer1', By Mohammad Ahmed Shah2, *Yıltan Bitirim1 2009 Fourth International Conference on Internet Monitoring and Protection IEEE .*

[13] 'A web-querying approach to Web People Search', By D. V. Kalashnikov, R. Nuray-Turan, and S. Mehrotra. Towards breaking the quality curse *In Proc. of Annual International ACM SIGIR Conference, Singapore, July 20–24 2008.*