2014

Ontology Based Personalized Search Engine

Divya Vemula
Grand Valley State University

Ontology Based Personalized Search Engine

By

Divya Vemula

August, 2014

A project submitted in partial fulfillment of the requirements for the degree of

Master of Science in

Computer Information Systems

Grand Valley State University

August, 2014

Jonathan Leidig **Date: 8/5/2014**

Table of Contents

1. Abstract:

An ontology is a representation of knowledge as hierarchies of concepts within domain, using a shared vocabulary to denote the types, properties and inter-relationships of those concepts [1][2]. Ontologies are often equated with classification of hierarchies of classes, class definitions, and the relations, but ontologies need not be limited to these forms. Ontologies are also not limited to conservative definitions, i.e., in the traditional logic sense that only introduce terminology and do not add any knowledge about the world (Enderton, 1972). To specify a conceptualization, axioms need to be proposed that constrain interpretation of defined terms [3].

Ontologies are frameworks for organizing information and are collections of URIs. It is a systematic arrangement of all important categories of objects and concepts within a particular field and relationship between them. Search engines are commonly used for information retrieval from web.

The ontology based personalized search engine (OPSE) captures the user's priorities in the form of concepts by mining through the data which has been previously clicked by them. Search results need to be provided according to user profile and user interest so that highly relevant search data is provided to the user. In order to do this, user profiles need to be maintained. Location information is important for searching data; OPSE needs to classify concepts into content concepts and location concepts. User locations (gathered during user registration) are used to supplement the location concepts in OPSE. Ontology based user profiles are used to organize user preferences and adapt personalized ranking function in order for relevant documents to be retrieved according to a suitable ranking. A client-server architecture is used for design of ontology based personalized search engine. The design involves in collecting and

storing client clickthrough data. Functionalities such as re-ranking and concept extraction can be performed at the server side of personalized search engine. As an additional requirement, we can address the privacy issue by restricting the information in the user profile exposed to the personalized mobile search engine server with some privacy parameters. The Prototype of OPSE will be developed on the web platform. Ontology based personalized search engines can significantly improve the precision of results.

2. Introduction:

Internet serves billion users with their information needs. Typically, users find the data either by searching or browsing. Search engines index billions of documents containing keywords. Faceted browsing is done by clicking through a hierarchy of concepts until the area of interest is found. The resulting node provides users with links of websites. Usually search and browse algorithms provide all users with same data. It is unlikely that all the user information needs are similar and one approach would not fit for all needs. In terms of searching, sometimes retrieved documents are reported to be irrelevant [11]. The major difficultly is that too much information is available, and keywords are not always appropriate to locate the information a user is interested in. Possibly, information retrieval will be more effective if a user's characteristics are taken into account. An effective personalization system would decide whether user is interested in a specific webpage and in the negative case, prevent it from being displayed on top. This means that ranking is performed based on user profiles. A major problem in searching data in search engines is the interactions between the users and search engines are limited by the small form factor. To return highly relevant results to the users, search engines should be able to profile the user's interests and personalize the search results according to the user's profiles.

3. Related Work:

A practical approach to capturing a user's interests for personalization is to analyze the user's clickthrough data. Leung et al., developed a search engine personalization method based on user's concept preferences and showed that it is more effective than methods that are based on page preferences [7]. Conversely, most of the previous work assumed that all concepts are of the same type. Detrimental to most commercial search engines is they return nearly the same results to all users. However, different users may require different information even for the same query. Many existing personalized web search systems are based clickthrough data to determine user's preferences. Joachim's proposed to mine document preferences from clickthrough data [5]. Later, W. Ng, L. Deng proposed to combine a spying technique together with a novel voting procedure to determine user preferences [6]. More recently, Leung et al., introduced an effective approach to predict users' conceptual preferences from clickthrough data for personalized query suggestions. Search queries are classified as non-geographical or location geo-based queries. Examples of location queries are "super markets at Baltimore", "Virginia historical places". Gan et al., developed a classifier to classify geo and non-geo queries [8]. It was found that a substantial number of queries were location queries focusing on location information. In order to handle the queries that focus on location information, a number of location-based search systems designed for location queries have been proposed. Yokoji et al., proposed a location-based search system for web documents. Location information was extracted from the web documents, which was converted into latitude-longitude pairs [9]. Later, Chen et al., studied on effective query processing in location-based search systems. A query is assigned with a query footprint that specifies the geographical area of interest to the user. Several algorithms are employed to rank the search results as a combination of a textual and a geographic score [10].

4. System Design:

In OPSE, client/server architecture clients are responsible for storing the user clickthrough, and ontologies are derived from the server. Tasks such as updating clickthrough and ontologies, creating feature vectors, and displaying re-ranked search results are handled by the clients. Ranking of the results are handled by the OPSE server. In order to reduce the data transmission between client and server, the OPSE client only needs to submit the query to the server; the server will return ranked search results according to the preference in the ontologies and user profile. The data transmission reduced as only the essential data (e.g., ontologies, query, search results) are transferred between client and server during the personalization process.

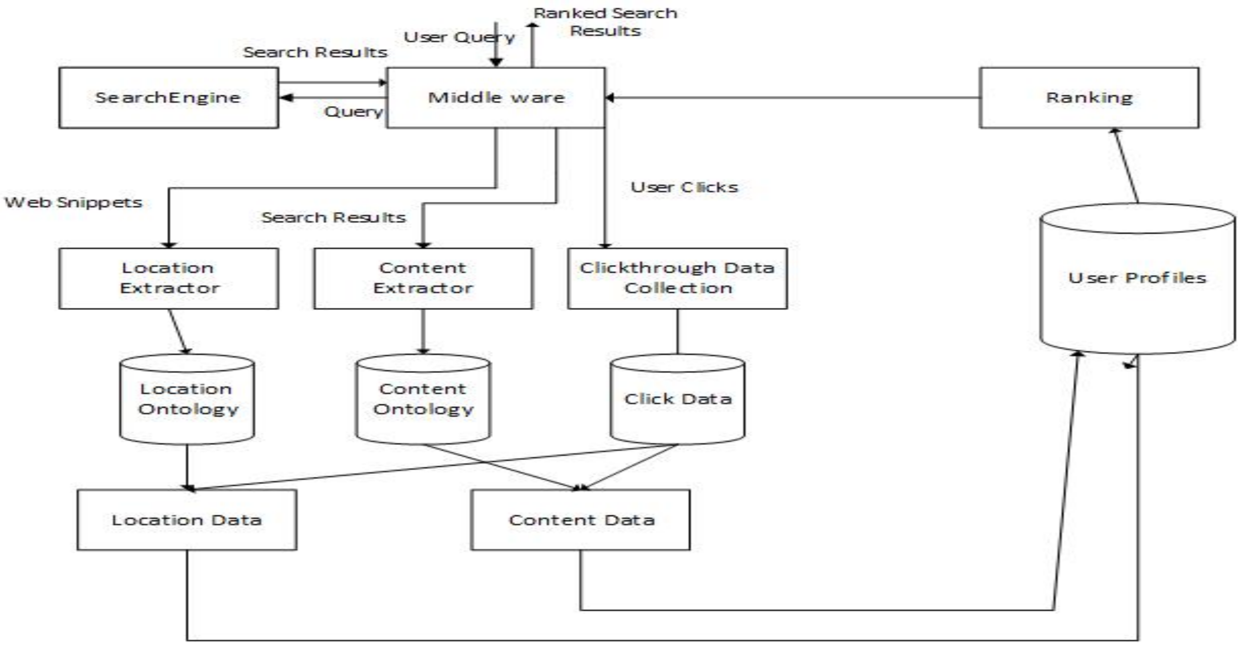


Figure 1: Architecture of OPSE

4.1 Profiling of user interests:

OPSE uses concepts to model preferences and interests of user. The concepts are further categorized into two different forms; content concepts and location concepts. The concepts are modeled as ontologies in order to capture the relationships between the concepts. Many observations say that the characteristics of the content concepts and location concepts are different. Two different techniques are used to build these forms (content ontology and location ontology). The ontologies indicate a possible concept space arising from a user's queries, which are maintained along with the clickthrough data for future preference adaptation. In OPSE, ontologies are used to model the concept space as they not only represent concepts but also capture the relationships between concepts. Content ontology and location ontology are mined and built from the search results.

4.1.1 Content Ontology

The interesting thing about content ontologies is that they represent both the available concepts as well as the user's historical interest in various concepts. For content concept all the keywords are extracted from the user query q. If a keyword exists in the web-snippets arising from the query q, it is treated as important concept related to q, as it coincides in proximity with the query in the top documents. The formula, which is inspired from problem of finding common item sets in data mining [12], is used to measure the importance of a keyword ci with respect to the query q:

$$support(c_i) = \frac{sf(c_i)}{n} \cdot |c_i|$$

Where *support(*c_i*)* is the frequency of the keyword phrase c_i, n is the number of web-snippets returned and $|c_{i|}$ is the number of terms in the keyword/phrase c_i. If the support of a keyword c_i is higher than the c_i is treated as a concept for q. Similarity and parent child relationship are the two propositions used to determine relationships between concepts for formulation of ontology.

Similarity: Coexisting concepts might represent same interest.

Parent-Child Relationship: Specific concepts often appear with general terms, while reverse is not true.

4.1.2 Location Ontology

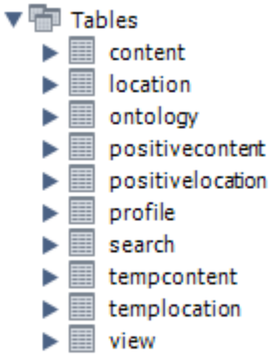
Concept of extracting location concept is different from content ontology. Location concepts are extracted from full documents, and it is difficult to extract similarity and parent child relationship from full documents because a limited amount of location concepts are present in the document. As all the locations are almost identified, it is possible to create an ontology by organizing all cities under their province or state, all provinces under their regions, and all regions under their country.

4.2 Diversity and location entropy:

To integrate user preferences in location and content ontologies into personalization we need to determine weights of content and location preference while integrating these concepts in search criteria. Adjusted weights for content and location preference are needed based on personalization. For a given query, if the content facet is more effective than location facet based on personalization more weight need to be given for content based preferences and vice versa.

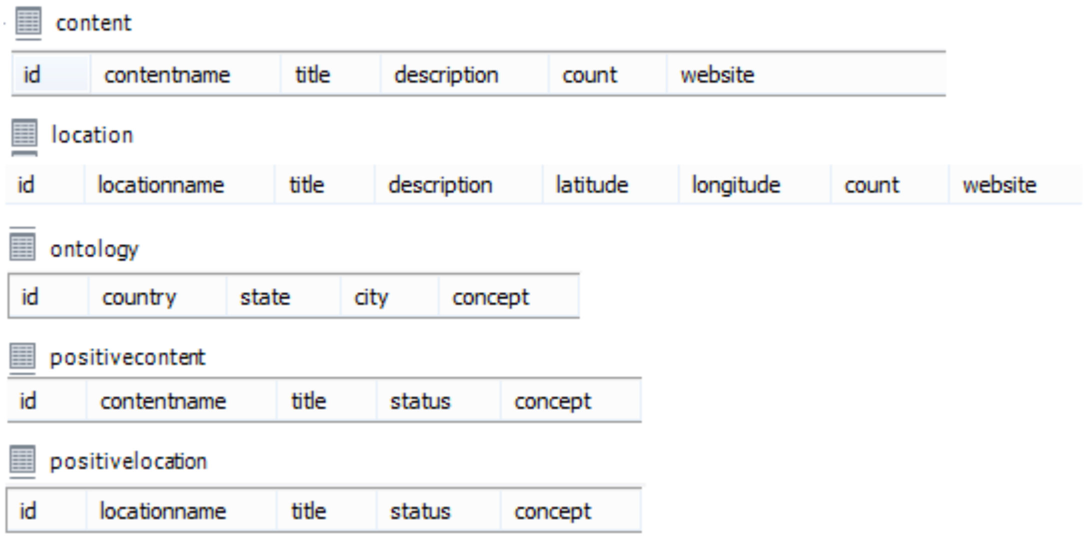
5. Implementation

MySQL: MySQL is an open source relational database management system. My SQL Workbench 6.1 was used as a visual database design tool for SQL development, administration and database design. For this application, a database was created with 8 tables including content, location, ontology, positivecontent, positivelocation, profile, search, tempcontent, templocation, and view.



Tables
- content
- location
- ontology
- positivecontent
- positivelocation
- profile
- search
- tempcontent
- templocation
- view

Figure 2: Database tables



content

id	contentname	title	description	count	website

location

id	locationname	title	description	latitude	longitude	count	website

ontology

id	country	state	city	concept

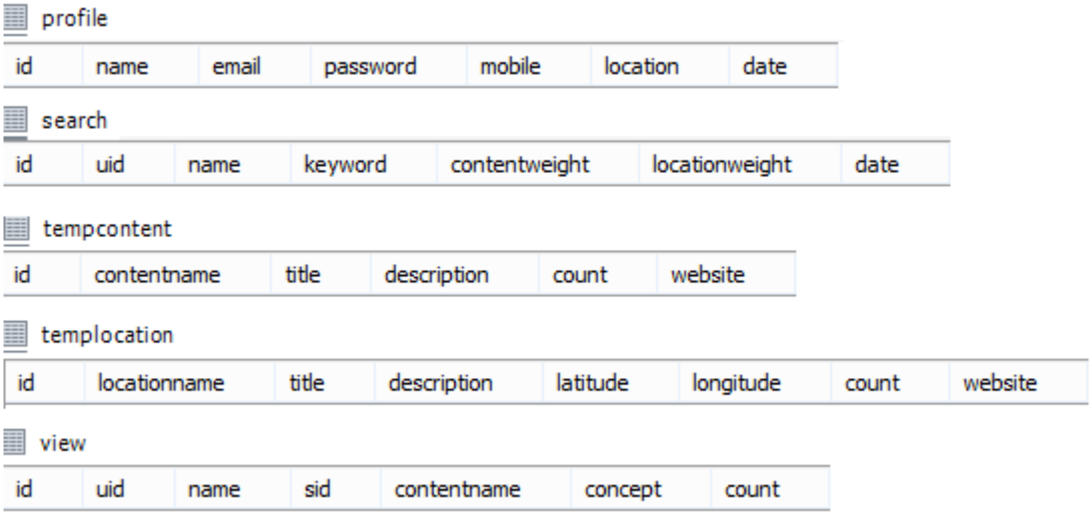
positivecontent

id	contentname	title	status	concept

positivelocation

id	locationname	title	status	concept

Figure 3 : Database Schema



profile

id	name	email	password	mobile	location	date

search

id	uid	name	keyword	contentweight	locationweight	date

tempcontent

id	contentname	title	description	count	website

templocation

id	locationname	title	description	latitude	longitude	count	website

view

id	uid	name	sid	contentname	concept	count

Figure 4: Database Schema

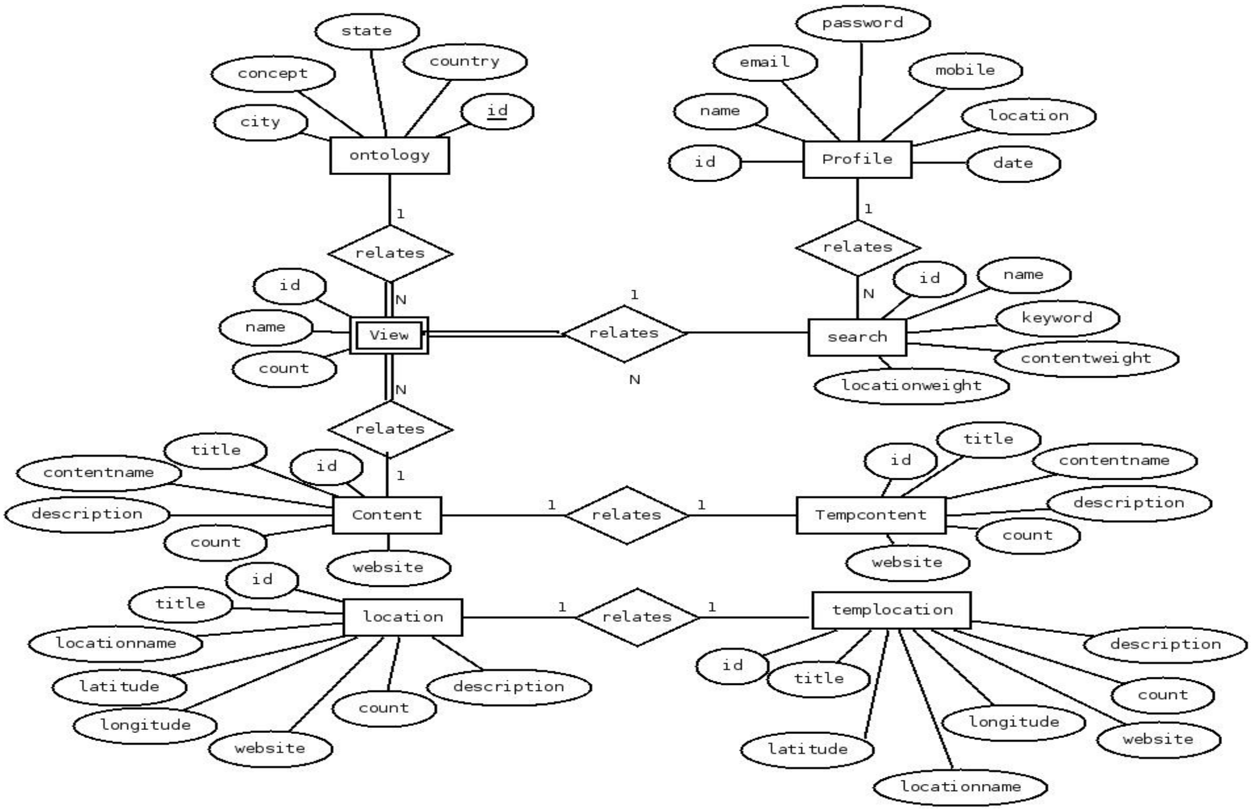


Figure 5: E-R Diagram

JavaScript:

JavaScript is a lightweight, interpreted programming language with object-oriented capabilities that allows you to build interactivity into otherwise static HTML pages. The general-purpose core of the language has been embedded in Netscape, Internet Explorer, and other web browsers. Advantages of java script are less server interaction, increased interactivity, and richer interfaces. JavaScript is used to provide user interface in prototype [13].

HTML:

HTML was developed with the intent of defining the structure of documents like headings, paragraphs, lists, and so forth to facilitate the sharing of scientific information between researchers. Now, HTML is being widely used to format web pages with the help of different tags available in HTML language[13].

CSS: CSS handles the look and feel part of a web page. CSS is used control the color of the text, the style of fonts, the spacing between paragraphs, how columns are sized and laid out, what background images or colors are used, as well as a variety of other effects.

JSP, Servlets and JDBC:

Java Server Pages (JSP) is a server-side programming technology that enables the creation of dynamic, platform-independent method for building Web-based applications. JSP have access to the entire family of Java APIs [13].

Java Servlets run on a Web or Application server and act as a middle layer between a request coming from a Web browser or other HTTP client and databases or applications on the HTTP server. Using Servlets, input is collected from users through web page forms, present records from a database or another source, and creates web pages dynamically [13].

JDBC is a Java API for database-independent connectivity between the Java programming language and databases. The JDBC library includes APIs for tasks commonly associated with database usage including connecting to data base, create SQL statements, executing SQL statements [13].

6. Conclusion and Future Work

Using Ontology based personalized search engine precision of results retrieved for a search query is improved with the help of user click through data and location. This lets us personalize search results for individuals. For future wok, to adapt to the user mobility GPS locations can be incorporated in the personalization process. Also privacy issues can be addressed controlling the amount of information exposed to the OPSE server.

References

1. Gruber, Thomas R. (June 1993). "A translation approach to portable ontology specifications". *Knowledge Acquisition,* 5 (2): 199–20. doi:10.1006/knac.1993.1008
2. Arvidsson, F. and Flycht-Eriksson, A. "Ontologies I". Retrieved 26 November 2008.
3. Ontologies and ontology engineering retrieved from http://www-ksl.stanford.edu/kst/what-is-an-ontology.html
4. E. Agichtein, E. Brill, and S. Dumais. "Improving Web Search Ranking by Incorporating User Behavior Information." Proc. 29th Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2006.
5. T. Joachim. "Optimizing Search Engines Using click through Data." Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, 2002.
6. W. Ng, L. Deng and D.L. Lee. "Mining User Preference Using Spy Voting for Search Engine Personalization," ACM Trans. Internet Technology, vol. 7, no. 4, article 19, 2007.
7. Q. Tan, X. Chai, W. Ng, and D. Lee, "Applying Co-Training to Click through Data for Search Engine Adaptation." Proc. Int'l Conf. Database Systems for Advanced Applications (DASFAA), 2004.
8. Q. Gan, J. Attenberg, A. Markowetz, and T. Suel . "Analysis of Geographic Queries in a Search Engine Log." Proc. First Int'l Workshop Location and the Web (LocWeb), 2008.
9. S. Yokoji. "Kokono Search: A Location Based Search Engine." Proc. Int'l Conf. World Wide Web (WWW), 2001.
10. Y.-Y. Chen, T. Suel, and A. Markowetz. "Efficient Query Processing in Geographic Web Search Engines." Proc. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR), 2006.
11. E. Casasola. "ProFusion PersonalAssistant: an agent for personalized information filtering on the WWW." Master's thesis, The University of Kansas, Lawrence, KS, 1998.

12. K.W. Church, W. Gale, P. Hanks, and D. Hindle. “Using Statistics in Lexical Analysis.” Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon, Psychology Press, 1991.
13. JSP,JDBC and Servlet codes referred from http://www.tutorialspoint.com
14. S. Gauch, J. Chaffee, and A. Pretschner“Ontology-Based Personalized Search and Browsing.” Web Intelligence and Agent Systems, vol. 1, nos. 3/4, pp. 219-234, 2003
15. J. Trajkova and S. Gauch. “Improving Ontology-Based User Profiles.” Proc. Conf. Recherche d’Information Assistee par Ordinateur (RIAO’04), pp. 380-389, 2004.
16. A. Sieg, B. Mobasher, and R. Burke. “Web Search Personalization with Ontological User Profiles.” Proc. 16th ACM Conf. Information and Knowledge Management (CIKM’07), pp. 525-534, 2007.
17. K. Sugiyama, K. Hatano, and M. Yoshikawa. “Adaptive Web Search Based on User Profile Constructed without any Effort from Users.” Proc. 13th Int’l Conf. World Wide Web (WWW’04), pp. 675-684, 2004.
18. J. Teevan, S.T. Dumais, and E. Horvitz. “Personalizing Search via Automated Analysis of Interests and Activities.” Proc. ACM SIGIR’05, pp. 449-456, 2005
19. D. Kelly and R.W. White. "A Study on the Effects of Personalization and Task Information on Implicit Feedback Performance.” in Proceedings of the 15th International Conference on Information and Knowledge Management, ACM, pp. 297-306, 2006.
20. *Ontologies and ontology engineering*. (n.d.). Retrieved from http://www.loa.istc.cnr.it/Ontologies.html