ISSN 2320 – 2602

Volume 3, No.7, July 2014

International Journal of Advances in Computer Science and Technology

Available Online at http://warse.org/pdfs/2014/ijacst02372014.pdf

An Optimized Smart Search Engine with meticulousness in Retrieval of Entertainment, Leisure and Travel related Informative Resources

KANIKA MINOCHA
IT DEPARTMENT, KJ SOMAIYA INSTITUTE OF MANAGEMENT STUDIES AND RESEARCH
RESEARCH SCHOLAR, JAGANNATHUNIVERSITY, JAIPUR
EMAIL:kanika.m@somaiya.edu

DISHAVERMA
DIAS, GURU GOBIND SINGH INDRAPRASTHA UNIVERSITY
RESEARCH SCHOLAR, JAGANNATH UNIVERSITY,JAIPUR
EMAIL:disha.verma.in@gmail.com

DR. BARJESH KOCHAR
DIRECTOR, SKITM
EMAIL:bjkochar@gmail.com

ABSTRACT

With more information being available on World Wide Web, General Search Engines are becoming ineffective for information access. Hence, the sheer volume of information presented by the General Search Engine is limiting its value. Generally, users have different information needs for his/her query. Search results should therefore be adapted to user information needs. Search Engines have become a primary tool for entertainment, leisure and travel planning. For these search engines to be successful, it is crucial, that information the user retrieves, should be relevant to the user and should be trustworthy as well. In this paper, we first propose a modular architecture of a Search Engine, focusing on integrating all scattered information related to entertainment, leisure and travel and then tailoring the information according to individual users. The proposed Search Engine will automatically acquire high quality facts about the various attractions, will filter them and will present only those pieces to user which is of interest to them. This Search Engine will be a combination of vertical and personalized search engine .We also quote some of the major differences between this search engine and general search engines. The focused set of websites being crawled for this category along with the application of various personalization approaches would lead to more efficient and accurate information retrieval.

Keywords: Vertical Search Engines, Personalization, Focused crawling, Entertainment Leisure and travel

1. INTRODUCTION

The development of the network technology has made internet the most invaluable source of information and World Wide Web the world's largest knowledge base. Web surfers obtain the desired information mainly by two ways, one is by travelling along the hyperlinks and the other is by using the search engines. Search engines begin to emerge in order to assist users to locate the much needed resources. Nevertheless, the distribution of information on Web, increasingly tends to be localized and specialized.

As the demand for theme information becomes strict, precision of retrieval becomes a crucial problem with general search engines. So general search engines ("GSE") such as Yahoo and Bing cannot perfectly meet the needs of all users as the theme gets submerged by these comprehensive type search engines. They often give irrelevant information to the user. Therefore, switching over from generalized to specialized searches has been a significant direction of development in the search engine market. Such search engines which focus on specialized topics are known as vertical search engines ("VSE"). VSE can be built more precisely and intelligently to meet the users need for information.

With the growing information on all facets of the tourism experience, general search engines such as Google and Yahoo! have become the "Hubble" of the Internet galaxy, enabling travelers to navigate through this space so as to find information that might be useful in the travel planning process [6].

As per a study done in [14] all the queries except the factual queries (what is) require personalization to some extent. But GSEs carry out all kind of searches by merely matching the keywords that the user inputs, and by using pagerank algorithm to sort the results, leading to low accuracy and recall ratio. VSEs can be made smarter and personalized by establishing a personalized model for the user. These search engines can establish the personalized model for the user according to user's location, income, age

, occupation, region, preference and other such factors. There can be two methods for establishing the user personalization model. First method is to let the user's reply certain questions and then to discover user's interest from his replies. The other method is heuristic discovery of the user interest according to the user's operation on the internet [3]. To illustrate the working and relevance of personalized VSEs, our paper would focus on a VSE based on entertainment, leisure and travel in India ("El&T") as a sub-case which can further be adopted in other domains with some tweaks and modifications. With focused set of websites being crawled for this category along with the application of various personalization approaches would lead to more efficient and accurate information retrieval [1]**.**

2. MOTIVATION

With the improvement in standard of living, the attitude of people towards lifestyle has changed. Nowadays, more and more Indians view entertainment as a part of healthy lifestyle and spend massive amount of money on entertainment and leisure travelling [2]. People are getting more and more used to searching of information related to food & drinks, movies, events and local attractions, intra and intercity excursions, transportation and accommodation, guidance, feedback and reviews about the abovementioned services, things to do, to and fro navigation guides etc.

On one hand with the increase in websites focusing on entertainment, leisure and travel, users have a wide range of information to select their actions and activities; on the other hand this wide range of online information poses a challenge for user to compare and choose. Although the web can give all needed information, the user often gets irrelevant information. Furthermore, due to enormous information on the web and because of spread of information in many different sources, users have to spend a lot of time in manually searching and integrating the information from those resources.

User usually visits more than a few websites and every time re-enters his query to find out the best deal before he makes the final purchase/decision. To ease out the process of comparison and selection, a personalized VSE for entertainment, leisure and travel can assist the users in making better decisions regarding such services.

3. RELATED WORK

The numerous research papers related to a search engine can be divided into two categories. The first category deals with the various approaches to personalization of search engines .The second category deals with the vertical search engines focusing on specialized searches

A lot of commercial systems are available in this area and the number of such systems is ever increasing. Few of them are commercial web systems like Bookmytrip.com, TripAdvisor.com etc.: These commercial systems do not integrate or comprehend all the information that is required for planning, nor do they adopt the results according the user without the effort of the user. In [8] there is research about applying web ontology language to the tourism sector In [4] a knowledge base which is based on ontology is designed and built using Protégé tool version 3.4.7. This ontology consists of information specific to tourism, attraction and cultural events in Bali.

In [5] there is a research on design of a crawler system in a vertical search engine, the crawler system that can extract professional information on the internet. In [7] there is focus on building profiles using search history whereas paper [11] focuses on user profile being specified explicitly. In [12] there is a study on how Open Directory Project (ODP) could be used to improve the ranking of search results. Each document gets categorized into one or more several nodes in the ODP hierarchy. In [10] there is an evaluation of personalizing strategies and it concludes that personalization may boost the performance of some of the queries while it may negatively affect the performance of queries that doesn't require personalizing

Based on our research very few papers have worked upon combining both vertical and personalization aspect of search engine to increase the efficiency and accuracy of information retrieval

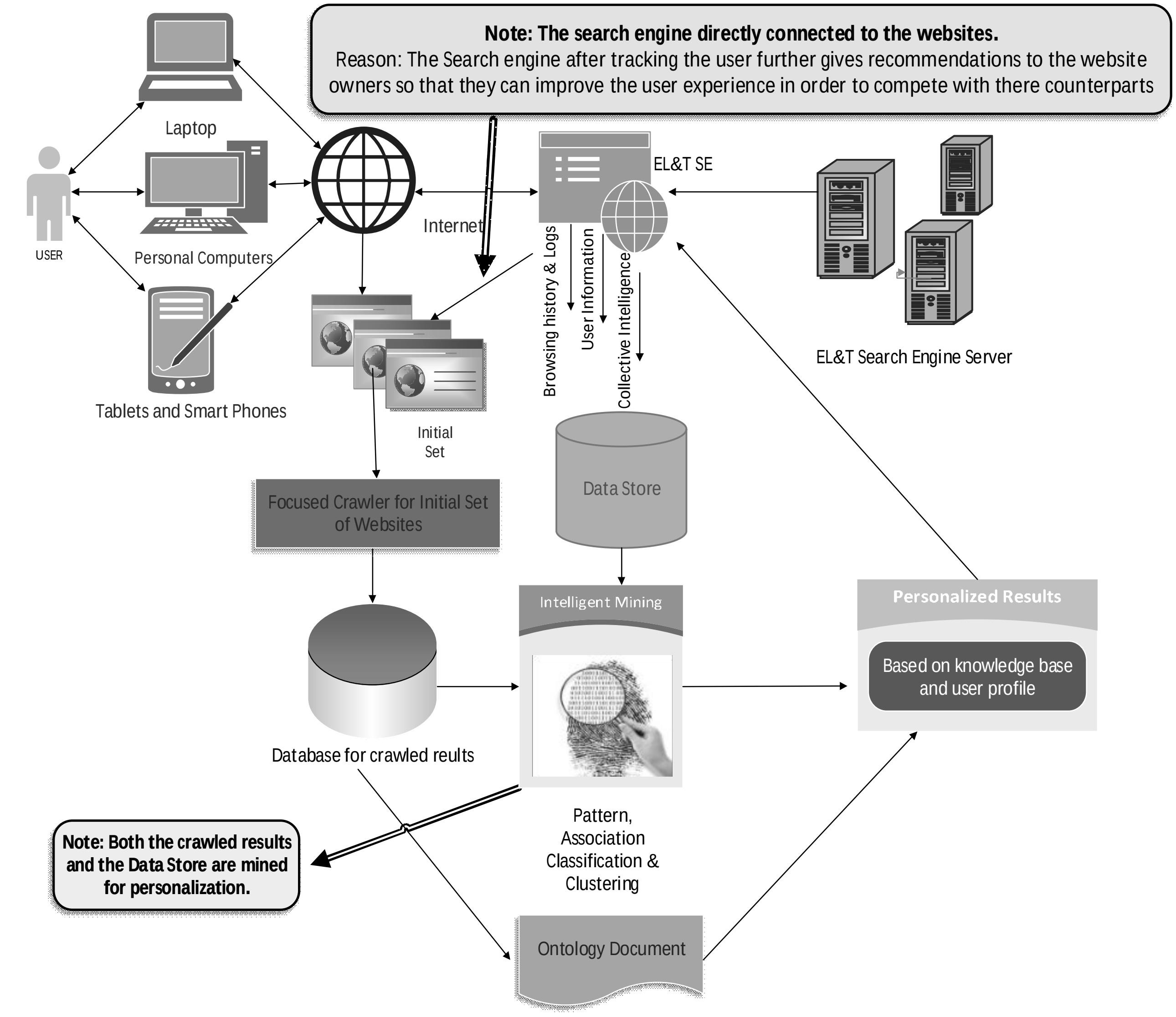
4. PROPOSED WORK

The EL&T SE will not require any special software. User only needs a personal computer, desktop or smart phone or tablet connected to internet. Figure 1 shows the complete framework for EL&T SE

.

4.1 About EL&T search engine

This search engine would focus on integrating all scattered information related to entertainment leisure and travel. It will have higher accuracy and recall ratio. It will also help the user in decision making, easy comparison and selection. It is capable of establishing the personalized model for the user according to user's location, age, occupation, preference and other such relevant factors. It can give recommendations to the website owners so that the websites can improve the user experience in order to compete with their counterparts



Note: The search engine directly connected to the websites.
Reason: The Search engine after tracking the user further gives recommendations to the website owners so that they can improve the user experience in order to compete with there counterparts

Note: Both the crawled results and the Data Store are mined for personalization.

Figure 1: The Framework for EL&T Search Engine

4.2 Initial Set

To collect the EL&T information, firstly all EL&T websites are identified and collected. Then the initial set is formed by extracting all sub-categories from these websites along with their link addresses. Figure2. shows few sample websites from which initial set of EL&T search engine will be formed.

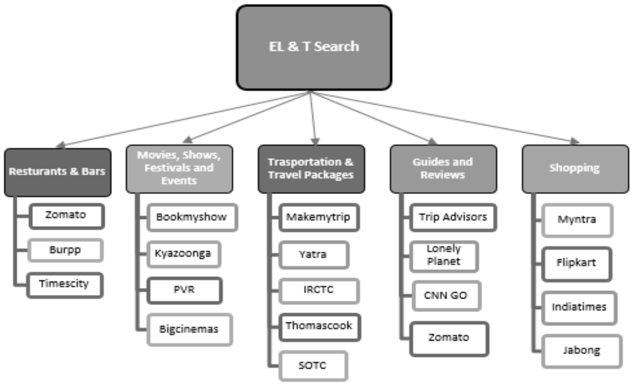


Figure2: Few Websites to be crawled for Each Category

4.3 Crawling Mechanism

Unlike regular crawling mechanism adopted by GSEs, building EL&T SE requires focused crawling. Focused crawler will only grab information from specific websites related to entertainment, leisure & travel (“**Initial Set**”) and will save the information in the database after preprocessing, filtration and analysis. GSE usually crawl from one page to the other following the links. On the contrary, EL&T linking addresses are available from the homepage of the websites. The requirement of eliminating duplicate information in EL&T SE is lower than in case of GSE.

4.4 Website update Notification System leading to Dynamic results

This system would provide a way for websites to supply data to search engines, as and when changes are made to their pages in terms of content or presentation. . It hosts software agents that would monitor the website for the changes and will notify the changes to the search engine.So the search engine would know when the index requires updating, and when can it avoid crawling those sites that have not changed since the last visit, leading to the crawler and indexer software being more efficient. It also keeps the index of search engine updated with the content that keeps on changing dynamically

4.5 Search Engine Application

When user will open the search engine application, he will be directed to the home page as shown in Figure 3. In order to realize the **personalized search**, the homepage of EL&T SE will have classification. The user can then inquire the category of his/her interest by clicking on the link which directs the user to the page chosen. Further to this, the user browsing history, user profile information, collective intelligence (user bookmarks of the webpages which are valuable to him/her etc.), server logs and cookie logs are saved in a data store. If any incorrect assumptions are made because of personalization user can explicitly specify the needed details using drag and drop menu.

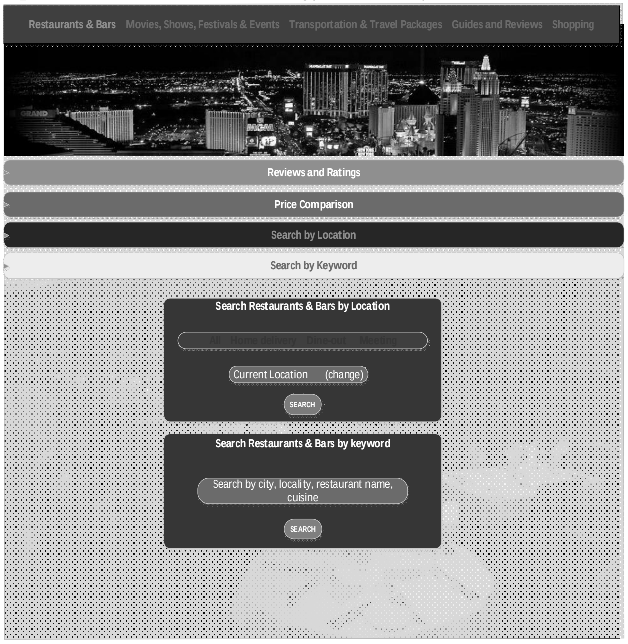

Restaurants & Bars
Movies, Shows, Festivals & Events
Transportation & Travel Packages
Guides and Reviews
Shopping
Reviews and Ratings
Price Comparison
Search by Location
Search by Keyword
Search Restaurants & Bars by Location
All
Home delivery
Dine-out
Meeting
Current Location
(change)
SEARCH
Search Restaurants & Bars by keyword
Search by city, locality, restaurant name, cuisine
SEARCH

Figure 3: Home Page for the search Engine

4.6 Intelligent Mining ,Ontology and Personalization

Rather than just retrieving what is already there in database, Ontology would allow the search engine to infer new information. The information gets arranged in a hierarchal manner that is much easier to extend as well as to query .This organization of information helps in deducing new facts from the already known ones. The ontology document allows filtering based on user own tastes and interests. Paper[14] describes such a travel ontology..

The crawled database storing the crawled documents from the initial set and the data store containing the user profile and preferences are then mined together intelligently to find patterns, clusters, associations[13] and to personalize the results according to the user.

Ontology, intelligent mining and cosine similarity would together pave the way for personalized results to the user. Through this mechanism EL&T SE will retrieve different results for different people at different time, different places etc.

4.7. Recommendations to the websites owners

Going one step further, EL&T SE will track the user and will be directly connected to the Initial Set of websites and will give recommendations to the website owners so that they can improve the user experience in order to compete with their counterparts. Additionally, by giving the recommendations to the website owners, EL&T SE can also generate revenue.

5. RESULTS AND DISCUSSIONS

Definition 1: The initial query provided by the user is $Q_{init =}$ {"Outdoor Recreation and activity ideas"}

Definition 2: The profile of the user is U_P that is built implicitly with the help of Browser cache, web logs and search logs

The profile is represented by categories related to ET&L. The profile is divided into categories using ODP. Initially demographic information like age, state, Date of birth, marital status etc: will be taken from the user. This demographic information can be used for personalization to a certain level in case the user is the first time user and no user profile exists

The categories are denoted by C= {c1, c2, c3,…....,c_n}
Each category is represented by keywords and their corresponding calculated weights.

c_i={($k_{ij,}$ w_{ij})}
where
$k_{ij=}$ keyword j in category i
$w_{ij=}$weight of keyword j in category i
where category could be Outdoor, Parties, Scuba Diving, Theme parks, Motorcycles ,Travel etc:

The Profile of the User U_1 is represented as shown in **Table1**
c1= Scuba Diving = {($k_{11=}$ "Wreck diving", w_{11}=0.8), ($k_{12=}$ "Underwater photography", w_{12} =0.5), ($k_{13=}$ "cave diving", w_{13} =0.7)}

Each area of interest in the user profile is represented as separate vector in order to provide a more accurate profile .Each keyword is represented using numerical weight that represents its importance within the profile vector. The numerical weight is calculated using tf-idf scheme.

Definition 3: Context of the user that includes Location L , Day D, Time T ,Weather W, Season S and Sentiment analysis.

The location along with time and other contexts will show the points of attention to the user around the current location along with the map and directions for the location. The time and day will be taken into consideration to find out that point of attention is open or closed, before suggesting it to the user. The sentiment analysis of the suggested location will also be done from the views represented by others on Facebook or twitter.

Scuba Diving			
Weight	0.8	0.5	0.7
Term	Wreck Driving	Underwater Photography	Cave Diving

Table1*: Profile of user U_1*

Definition 4: Initial set of default websites to be crawled I_s= {Zomato,Timescity, Burpp, Bookmyshow, PVR, ………………TripAdvisor}. These initial set of websites will be crawled and the results will be filtered by the information filtering agent using user query for finding out recreational activities.

Let this be one of the document d_1

d_1= WLA chapters provide opportunities for friends, families, and neighbors to participate in a wide range of **outdoor recreational activities** including **wreck diving**. So use the chapter's property to expose not only the card-carrying members but the public to the great outdoors. Teach them **outdoor recreational skills**, and you'll instill a conservation ethic and recruit new members and supporters at the same time

Query q_1= "Outdoor Recreation and activity ideas"

Let the **document vector** be A_i and the **query vector** B_i as shown in **Table 2**

$$\text{similarity} = \cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

$$\text{Cos}(\Theta)= \frac{(2.1+2.1+2.1+0.1)}{(\sqrt{2^2+2^2+2^2+0^2})(\sqrt{1^2+1^2+1^2+1^2})} = 0.86 \text{ approximately}$$

Table 3 shows the range of cosine similarity

−1 meaning exactly opposite
1 meaning exactly the same
0 usually indicating independence
In-between values indicating intermediate similarity or dissimilarity.

Table 3 : Range of the Similarity result of cosine similarity

The results will be initially ranked (I_R) according to the cosine similarity of the pages. After that the rank is further filtered altered according to the weight of the keywords in the profile using the equation

***Final rank*=β*keyword weight + (1- β)*I_R**

Where β has a value between 0 and 1. When β has a value of 0 purely initial rank is taken into consideration where user profile will play no role. If β has a value of 1 then only user profile will decide the order of results. When β =0.5. Both I_R and user profile will decide the rank of the document.

Let k_{11} and k_{12} be the keywords in the user profile that matches the text in document retrieved then

keyword weight =k_{11}+k_{12}

If there are no keywords that matches the text in document then keyword weight=0. Since in the above document d_1 the keywords wreck diving are there so

***Final rank*=0.5*0.8+O.5*0.86=0.83**

	Outdoor	Recreation	Activity	ideas
A_i	2	2	2	0
B_i	1	1	1	1

Table 2: The number of times each word appears in document and query

6. COMPARISON OF EL&T WITH GENERAL SEARCH ENGINES

Parameter	GSE	EL&T
Retrieval source	Entire Internet	Websites related to entertainment(Initial Set)
Search	Less Dynamic	Search related to EL&T appears to be dynamic because of factors like, travel information being seasonal, difference in costing on weekend and weekdays, discount and festive offers being offered on food & drinks and other events, change in ratings of restaurants, etc.
Personalized	Same for all users	Personalize model as user's location, age, occupation, region, time, weather, season sentiment analysis and other such factors.
Authentic and Trustworthy	Results from All resources	Results from only authentic and trustworthy resources.

Table 4: Comparing EL&T with GSE

7. CONCLUSION AND FUTURE WORK

With the above methodology we can achieve an optimized smart search engine which has precision of retrieval, is dynamic in nature and shows results which are authentic and trustworthy .It will act as one stop portal for the users looking for accurate and precise information on entertainment, leisure and travel. Therefore, this search engine will meet the needs of the users as comprehensively as possible. This will also help the website owners to provide the finest user experience possible based on the feedback received from the EL&T SE and which will ultimately result in healthy competition within the market.In future more advanced techniques for improving the user profiles can be adopted and integrated with the architecture so as to cater to both long term and short term interests of the user hence suggesting both permanent and temporary point of interests.

REFERENCES

1. Meng cui; songyun hu "search engine optimization research for website promotion", *information technology, computer engineering and management sciences (icm), 2011 international conference on*, on page(s): 100 - 103 volume: 4, 24-25 sept. 2011.

2. Longyan luo; yong wang, "status and development strategies of chinese travel search engines," *information technology and artificial intelligence conference (itaic), 2011 6th ieee joint international* , vol.2, no., pp.416,419, 20-22 aug. 2011.

3. Jun gong, "analysis the idea of personalized search engine based on user behavior," *computer application and system modeling (iccasm), 2010 international conference on* , vol.5, no., pp.v5-450,v5-452, 22-24 oct. 2010.

4. Kuntarto, g.p.; gunawan, d., "dwipa search engine: when e-tourism meets the semantic web," *advanced computer science and information systems (icacsis), 2012 international conference on* , vol., no., pp.155,160, 1-2 dec. 2012.

5. Min li; jun zhao; tinglei huang, "research and design of the crawler system in a vertical search engine," *intelligent computing and integrated systems (iciss), 2010 international conference on* , vol., no., pp.790,792, 22-24 oct. 2010.

6. Xiang zheng;panbing;law rob," an analysis of search engine use for travel planning",", information and communication technologies in tourism,2010,pp 381-392.

7. Speretta, m., &gauch, s. Personalized search based on user search histories. In web intelligence, 2005. Proceedings. The 2005 ieee/wic/acm international conference on (pp. 622-628). Ieee. September (2005).

8. Soza a. M. And garrido l. C. T.v. 2010. *Web ontology language applied to the tourism sector*. In prospect. Vol. 8,no. 1, enero - junio de 2010, págs. 87-93.

9. Dou z., song r., and wen j. A large-scale evaluation andanalysis of personalized search strategies. In proc. 33rd annualint. Acm sigir conf. On research and development in information retrieval, 2007.

10. Bennett, p. N., svore, k., &dumais, s. T.classification-enhanced ranking. In *proceedings of the 19th international conference on world wide web* (pp. 111-120). Acm. April, (2010).

11. Chirita, p. A., nejdl, w., paiu, r., &kohlschütter, c. ,using odp metadata to personalize search. In proceedings of the 28th annual international acm sigir conference on research and development in information retrieval (pp. 178-185). Acm, august (2005).

12. Pierrakosdimitrios;paliourasgeorgios , papatheodorouchristos , constantine d. Spyropoulos, web usage mining as a tool for personalization: a survey, user modeling and user-adapted interaction, v.13 n.4, p.311-372, november 2003 [doi>10.1023/a:1026238916441].

13. F. Shi and j. Li, "toward integration travel information data using information extraction and instance matching," in proceedings of the websci'09: society on-line, 2009.

14. Justin briggs,http://justinbriggs.org/better-understanding-personalized-search.