ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Personalized Mobile search engine with GPCA Algorithm

K. Murugan, S. Suma

PG Student, M.E (CSE), Valliammai Engineering College, Chennai, Tamilnadu, India[1]

Asst. Professor, Valliammai Engineering College, Chennai, Tamilnadu, India[2]

ABSTRACT - We proposed a search engine that captures clickthrough data with time preferences which improves personalization more effective. User behavior is collected with content (user query), location (using GPS) and time of query. Timing information is collected so that user preference as per time is collected from that behavior of user is well known. GPCA algorithm is used for time information. Normally query result will be web popular result will present in top 100 results but user requested will be not get so user clicked data are stored as user preference and for future query result will be reranked with location, content and time. Ersonalization become more effective due to time in user behavior which makes system to understand what will be user requested and make ranking of resulted web page.

KEYWORDS: GPCA, subspace, data points, clicked

I. INTRODUCTION

Today globally all having mobile phones which will be connected to network and GPS present in it helps to get user preferred output is ranked using PMSE[1] which uses location ontology and content ontology which will rank the web pages according nearby location with content requested by user. User personalization is improved by collecting time of user is collected with its content and location of user query.

II. RELATED WORKS

1."PMSE: A Personalized Mobile Search Engine[1]" Kenneth Wai-Ting Leung, DikLun Lee, Wang-Chien Lee proposed a system that captures user preferences in form of user clicked data. Location information is easily obtained by using mobile GPS system. PMSE divides concept into location concept and content concept.
Advantage: 1.Ontology which make user query to simplified one. 2. Location ontology helps to get user preference in terms of new locationDisadvantage: 1. Sequential queries are not possible in this system. 2. One column per aggregation only achieved in this system.

2."Improving web search ranking by incorporating user behavior information,[7]" E. Agichtein, E. Brill, and S. Dumais proposed a system which incorporates user behavior by ordering tope results in real web search engine setting. In this paper user feedback incorporated into re-ranking and user feedback is compared with other web results.
Advantage: 1.Implict feedback is inserted into search process. 2. Implicit feedback is directly used in training the ranking function.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

III. PERSONALISED MOBILE SEARCH ENGINE

A practical approach to capturing a user's interests for personalization is to analyze the user's click through data [5], [10], [15]. Search engine personalization method based on users' concept preferences and showed that it is more effective than methods that are based on page preferences [12]. However, most of the previous work assumed that all concepts are of the same type. Observing the need for different types of concepts, we present in this paper a personalized mobile search engine PMSE[1], which represents different types of concepts in different ontologies. In particular, recognizing the importance of location information in mobile search, we separate concepts into location concepts and content concepts. For example if user new to our college gives a query "hotel" and click on the search result about the hotels around our college. Accordingly, PMSE[1] will favor results that are concerned with hotel information in our college for future queries on "hotel". The introduction of location preferences offers PMSE[1] an additional dimension for capturing a user's interest and an opportunity to enhance search quality for users.

In PMSE[1] by adopting the metasearch approach which replies using Google to perform an actual search. The client is responsible for receiving the user's requests, submitting the requests to the PMSE[1] server, displaying the returned results, and collecting his/her clickthroughs in order to derive his/her personal preferences. The PMSE[1] server, on the other hand, is responsible for handling heavy tasks such as forwarding the requests to a commercial search engine, as well as training and reranking of search results before they are returned to the client. The user profiles for specific users are stored on the PMSE [1] clients.

A. CONTENT ONTOLOGY

Our content concept extraction method first extracts all the keywords and phrases (excluding the stop words) from the web-snippets2 arising from query. If a keyword/phrase exists frequently in the web-snippets arising from the query, we would treat it as an important concept related to the query, as it co-exists in close proximity with the query in the top documents. The following two propositions to determine the relationships between concepts for ontology:

a)Similarity: Two concepts which coexist a lot on the search results might represent the same topical interest for user query.
b) Parent-Child Relationship: More specific concepts often appear with general terms, while the reverse is not true.

B. LOCATION ONTOLOGY

Our approach for extracting location concepts is different from that for extracting content concepts. We observe two important issues in location ontology formulation. First, a document usually embodies only a few location concepts, and thus only very few of them co-occur with the query terms in web-snippets.
To alleviate this problem, we extract location concepts from the full documents. Second, the similarity and parent-child relationship cannot be accurately derived statistically because the limited number of location concepts only available.

C. CONCEPT ENTROPY

In order to seamlessly integrate the preferences in these content and location facets into one coherent personalization framework, an important issue we have to address is how to weigh the content preference and location preference in the integration step. To address this issue, we propose to adjust the weights of content preference and location preference based on their effectiveness in the personalization process. For a given query issued by a particular user, if the personalization based on preferences from the content facet is more effective than based on the preferences from the location facets, more weight should be put on the content-based preferences; and vice versa. In information theory [17],

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

entropy indicates the uncertainty associated with the information content of a message from the receiver's point of view. The notion of personalization effectiveness is derived based on the diversity of the content and location information in the search results as content entropy and location entropy.

D. PERSONALIZATION EFFECTIVENESS

From content entropy and location entropy, a query result set with high content/location entropy indicates that it has a high degree of ambiguity. Thus, applying personalization on the search results helps the user to find out the relevant information. On the other hand, when the content/location entropy is low, meaning that the returned result set is already very focused and should have matched the query quite precisely, personalization can do very little in further improving the precision of the result. For click entropies, we expect that the higher the click content/location entropies, the worse the personalization effectiveness, because high click content/location entropies indicate that the user is clicking on the search results with high uncertainty, meaning that the user is interested in a diversity of information in the search results.

IV. SUBSPACE ANALYSIS

In our system personalization is achieved by capturing the click-through data with content ontology and location ontology. In which user personalization is achieved to extent where user behavior is not calculated with time. For example if user requested for a Shop in end of year, now the user clicked data are stored as personalization and when user searching on "New Year" for discount sale user will get result as per previous result. Therefore time plays important role in calculating user behavior. To calculate we go for subspace analysis.

From the result of search engine user clicked data are collected. From these clicked data we get information about user identity, user request, time of request and link clicked by user. Clicked data provides two information semantics events and time of query issued. GPCA involves four steps:1.Polar transformation, 2.Subspace estimation, 3.Subspace pruning, 4. Cluster generation.

a) Polar transformation:

Clicked data collected are converted to 2D polar space. Each query entered and corresponding clicked pages is mapped to a point in polar space such that angle θ and radius r of the point respectively which resembles the semantics and the time of event occurred illustrated in Fig 1.

b) Subspace pruning:

Data points present after subspace estimation noise only removed, some unwanted data points are also present therefore pruning is achieved. In Polar transformation method temporal burst and semantic burst of query sessions are reflected some are along subspace direction and orthogonal direction of subspace. to estimate certainty of distribution of data points along both direction, we project the data points to respective directions and calculate the respective histograms of data points in directions.

c) Cluster generation:

After pruning uninteresting subspaces, events can be detected from the remaining subspaces by clustering. Particularly, we detect various events from interesting subspaces by employing a non-parametric clustering method called Mean Shift [2]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Algorithm:

Input : {S} set of query sessions $\{S_1,S_2,\ldots S_n\}$

{f} set of principle components $\{f_1,f_2,\ldots,f_n\}$;

{H} set of histograms;

T_i time of session;

ζ threshold;

D data points

Output: set of detected events by mean shift clustering.

Steps

BEGIN

//Transform query sessions to polar space

foreach session $S_i \in \{S\}$

Compute radius r_i

$$r_i = \frac{T(S_i) - \min(T(S_j))}{\max_j (T(S_i) - \min(T(S_j)))}$$

where $T(S_i)$ is the occurring time of query sessions S_i

$r_i$takes value in the range {0,1}

//semantic similarity between two query sessions

let $S_1 = (Q_1, P_1)$ and $S_2 = (Q_2, P_2)$ as

$$sim(S1,S2) = \alpha \times \frac{|Q1 \cap Q2|}{\max\{|Q1|,|Q2|\}} + \frac{(1-\alpha)|P1 \cap P2|}{\max\{|P1|,|P2|\}}$$

for each principle component $f_i \in \{f\}$

calculate

$$\theta_i = \frac{f_i - \min_j(f_i)}{\max_j(f_i) - \min_j(f_i)} \times \frac{\pi}{2}$$

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

θi is restricted to [0, π/2].

//Estimate subspaces of query sessions

For given data point x_i, weight w_i assigns

$$W_i = \frac{1}{1 + (s(NN_{xi}) + n(NN_{xi})) \times \frac{s(NN_{xi})}{s(NN_{xi})}}$$

where$s(NN_{xi})$ is the variance of xi's K nearest neighbors along the subspace direction and $n(NN_{xi})$ is the variance of its neighbors along the orthogonal direction of the subspace.
For all $x_i \in D$ && $h_i \in \{H\}$;

//SubspacePruning

Calculate

$I(s_i) = 1 - [-p\sum_{i=1}^{n} h_i \log h_i - (1-p)\sum_{i=1}^{n} v_i \log v_i]$

//Cluster generation

If $(I(S_i) < \zeta)$

Drop the data items

else

Event list (mean_shift(event_detect=meanshift(D))

END

Table1. User Profile In Existing System

User	Query	Location	Content
45	Motels	Potheri	Chinese, Contential
87	Temple	MM nagar	Ayyappan
15	Hotels	Spkoil	Room rate, A/C rooms
47	Parks	Tambaram	Park
58	Shopping mall	Vandular	Mall,shopping center

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

Table 2. User Profile In Proposed System

User	Query	Location	Content	Time
45	Motels	Potheri	Chinese, contential	14/03/12
87	Temple	MM nagar	Ayyappan	12/01/11
15	Hotels	Kattankulthur	Roomrate, A/C rooms	22/12/10
47	Parks	Tambaram	Park	04/08/03
58	Mall	Vandular	Mall, Shopping mall	05/01/12

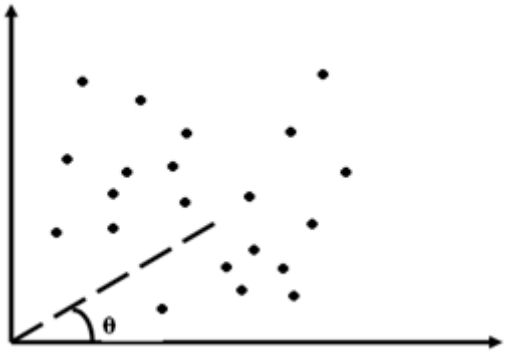


Fig1. Polar Representation

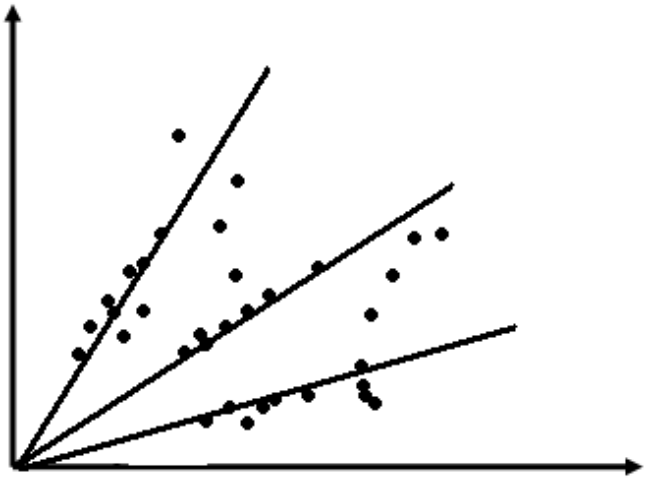


Fig2.Subspace Estimation

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

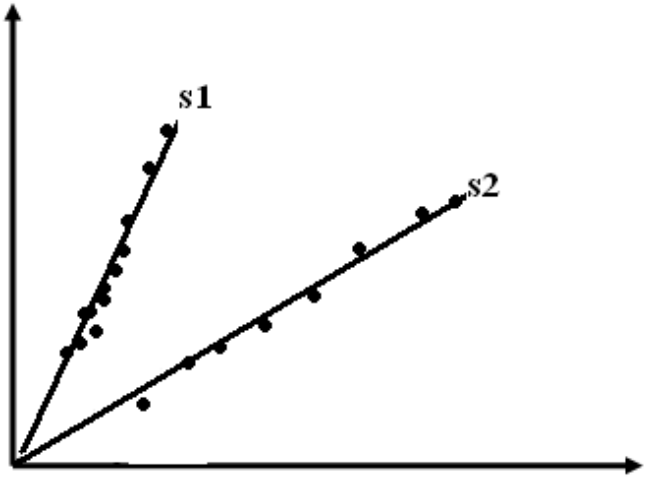
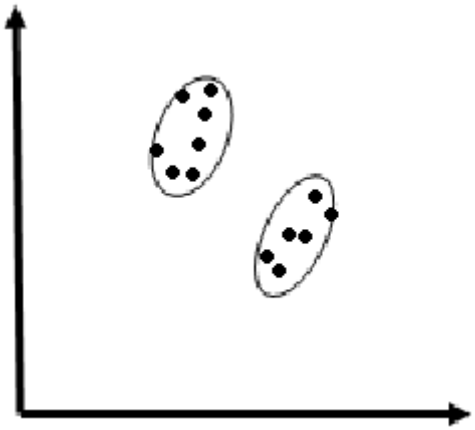


Fig3.Subspace Pruning



V. CONCLUSION

Thus our search engine with time space makes personalization effective. Location ontology helps to get content nearby the user and rank them by clickedthrough data. GPCA algorithms make mining with time model and make personalization with query.

REFERENCES

[1]PMSE: A Personalized Mobile Search EngineKennethWai-Ting Leung, DikLun Lee, Wang-Chien Lee. In IEEE,2013.

[2]D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. In IEEE TPAMI, volume 24, 2002.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization) Vol.2, Special Issue 1, March 2014

Proceedings of International Conference On Global Innovations In Computing Technology (ICGICT'14)

Organized by

Department of CSE, JayShriram Group of Institutions, Tirupur, Tamilnadu, India on 6th & 7th March 2014

[3] W.-S. Li, K. S. Candan, Q. Vu, and D. Agrawal. Retrieving and organizing Web pages by "information unit". In WWW, 2001.

[4] Z. Li, B. Wang, M. Li, and W.-Y. Ma.A probabilistic model for retrospective news event detection.In SIGIR, 2005.Mean shift: A robust approach toward feature space analysis", DorinComaniciu, Peter Meer.

[5] G. Pass, A. Chowdhury, and C. Torgeson. A picture of search.In The First International Conference on Scalable Information Systems, 2006.

[6] DECK: Detecting Events from Web Click-through Data Ling Chen, Yiqun Hu, Wolfgang Nejdl. IEEE 2008.

[7]E. Agichtein, E. Brill, and S. Dumais, "Improving web search ranking by ncorporating user behavior information," in Proc. of ACMSIGIRConference,2006.

[8] E. Agichtein, E. Brill, S. Dumais, and R. Ragno, "Learning user interaction models for predicting web search result preference" in Proc. of ACM SIGIR Conference, 2006.

[9] Y.-Y. Chen, T. Suel, and A. Markowetz, "Efficient query processing in geographic web search engines," in Proc. of ACM SIGIR Conferecn, 2006.

[10] 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, 1991.

[11] Q. Gan, J. Attenberg, A. Markowetz, and T. Suel, "Analysis of geographic queries in a search engine log," in Proc. of LocWeb Workshop, 2008.

[12] T. Joachims, "Optimizing search engines usingclickthrough data," in Proc. of ACM SIGKDD Conference, 2002.

[13] K. W.-T. Leung, D. L. Lee, and W.-C. Lee, "Personalized web search with location preferences," in Proc. of IEEE ICDE Conference,2010.

[14] K. W.-T. Leung, W. Ng, and D. L. Lee, "Personalized concept-based clustering of search engine queries," IEEE TKDE, 2008.

[15] "ENHANCED TRUSTWORTHY AND HIGH-QUALITY INFORMATION RETRIEVAL SYSTEM FOR WEB SEARCH ENGINE" Sumalatha Ramachandran, SujayaPaulraj, Sharon Joseph and VetriselviRamaraj. IEEE 2009.

[16] "Personalized Web Search by Using Learned UserProfiles in Re-ranking" Jia Hu and Philip K. Chan. IJCSI 2009.