ISSN: 2277–3061 (online)
International Journal of Computers & Technology
Volume 2 No. 3, June, 2012

A Context Model For Focused Web Search

Sushil Kumar
Assistant Professor(CE)
YMCA University Of Science and Technology
Faridabad(Haryana)

Naresh Chauhan
Professor(CE)
YMCA University Of Science and Technology
Faridabad(Haryana)

ABSTRACT

In the existing web search systems, the information retrieval is performed using a single query and mapping it to a set of documents. From a single query, however, the search systems can only have very limited clue about the user's information need. The user's context and his environment are ignored while searching the information resulting in irrelevant search results. These irrelevant search results increase the cognitive overhead of the user in filtering them out and finding useful information. Therefore, the search systems must incorporate context information regarding user and his environment search the highly relevant web pages. This paper prepares an Entity-Centric model for the context and proposes a framework for context-aware focused web search system that considers the various context features and returns highly relevant search results to the user.

General Terms

Web search, web crawling.

Keywords

Context model, context web search, entity-centric context model

1. INTRODUCTION

The existing web search systems perform retrieval decisions based solely on the single query and document collection. Typically, web users are expected to express their need via a set of query terms submitted to the search system and the query is then compared to each document in the collection resulting in a set of potentially relevant documents consisting of irrelevant also. Thus, it forms a query-document matching function as shown in Fig.1. However, a few results returned may be valuable to the user. It means that web search system has ignored the information about the actual user and the search context and returns many irrelevant results along with the relevant results. It increases the cognitive overhead of the user as he puts additional effort and concentration necessary to maintain several tasks (such as scanning the pages, following several links, making decisions about which links to follow and which to abandon, etc.) at one time. It was observed [12] that 74% users realize that they loose their concentration towards the original information they were seeking and become frustrated after lot of cognitive overhead.

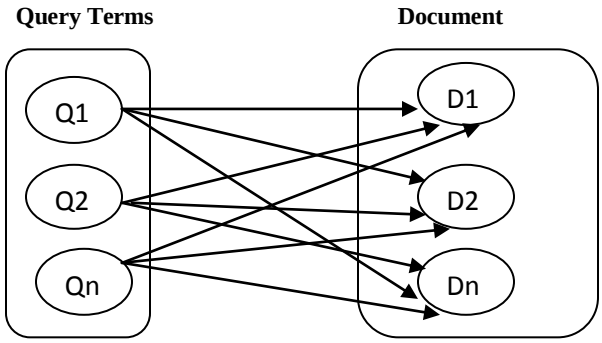

Fig.1 Query-Document Function

The search quality of web pages can be improved by focused web search [2,3,13-15], which aims to search and retrieve only the subset of the www that pertains to a specific topic of relevance. The existing focused systems [3,4] adopt different strategies for computing the words' frequency in the web documents. If higher frequency words match with the topic keyword, then the document is considered to be relevant and it is downloaded without regards to the context of the keywords leading to non-goal oriented search. In fact, a context is a very important component towards building a goal-oriented web search system. In a workshop held [5] at the Center for Intelligent Information Retrieval, University of Massachusetts Amherst, September 2002, the contextual retrieval was considered as the long-term challenge in the information retrieval and was defined as:

Combine search technologies and knowledge about query and user context into a single framework in order to provide the most "appropriate" answer for a user's information needs.

To identify the importance of context in focused web search, consider the following examples when the user enters his query keywords to the keyword-driven search engines:

- If the user enters the query as 'mouse', then it has various interpretations. To a computer professional,

it is the name of an input device, whereas to a layman it is simply name of an animal.

- If the user enters the word 'Java', then the user wants to learn about the programming language or wants the information about the an island 'Java' in Indonesia.
- If a user enters the name of a product only as a search keyword, then it is not clear whether he is looking for a company that sells this product or technical details about the product.

The above examples indicate that a general user is not aware about providing the context of the information what he desires. The user provides random combination of keywords to obtain the information. Thus, it becomes a challenge how to provide the correct context concerned with the user query. More specifically, the following problems are critical in providing the context:

- **Incomplete query:** The novice or inexperienced users can have difficulty in formulating the queries. They usually do not provide complete descriptions about their information needs. Most of the users are reluctant to enter multiple keywords in the query and do not enter more than two words [12,1]. Consequently, they get the results that are not matching with their requirements.
- **Unaware about the context keywords:** Most of the time, user do not enter the related keywords of the area in which they are searching the information. 27.8 users never know the proper keywords of the information they seek. 28.3 users sometime know the keywords and 31.8 users know them seldom [12].
- **Noisy Context**: If the user tries to reform his previous query, then it may be possible again that he adds some random keywords that negatively affects the retrieval performance leading to unrelated information.

Hence, from a single query without proper context, the search system can only have very limited clue about the user's information need and consequently download the irrelevant search results as shown in Fig. 2.

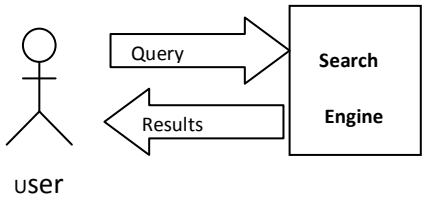

Fig. 2 Conventional method for retrieving information based on only the query terms supplied by the user

This approach of retrieving the information is known as 'Black box'. In this approach the information retrieval systems are static and not aware of context in which they operate. Therefore, for any query, it is assumed that relevant information will be found through the predetermined set of documents stored in the database. Thus, the topical relevance is not the only issue for focused web search, but context relevance should also be considered. If the user issues one keyword, then its relevant context must also be known. In fact, the contextual information with the query terms must be merged to improve the quality of search results as shown in Fig 3 and query-document function should be modified to incorporate the contextual terms also (see Fig. 4). The contextual information incorporated with the original query terms, thus, optimizes the information closer to the user demand.

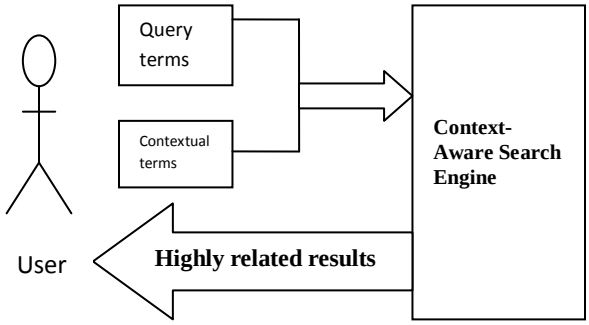

Fig. 3 Context driven web search system

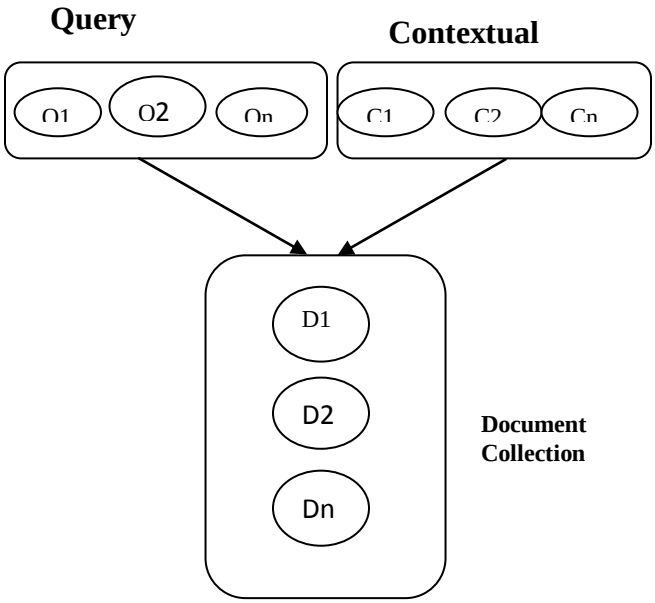

Fig.4 Context-Query-Document Function

In this way, the ability of a search system to respond to the contextual information allows to learn and predict what information a searcher needs, decide how and when this information should be presented and distinguish between different search goals and user preferences.

2. Context Typology

A categorization of context types will help the designers of context-aware web search system uncover the most important parts of context. Some of the types are described below:

- **Primary and Secondary Context:** Identity, location, activity and time are the primary context types for characterizing the situation of a particular entity. Identity refers to the ability to assign a unique identifier to an entity. Location includes orientation and elevation, as well as all information that can be used to deduce spatial relationships between entities, such as co-location, proximity, or containment. Status or activity identifies intrinsic characteristics of the entity that can be sensed. Time is most often used in conjunction with other pieces of context, either as a timestamp or as a time span, indicating an instant or period during which some other contextual information is known or relevant.

  The primary context types not only answer the question of who, when, and where, but also act as indices into other sources of contextual information. For example, given a person's identity, we can acquire many pieces of related information such as phone numbers, addresses, relationship to other people in environment, etc. It indicates that the primary pieces of context for one entity can be used as indices to find secondary context for that same entity as well as primary context for other related entities [6].

- **Short and Long term Context:** Short-term context is the immediate surrounding information which throws light on a user's current information need in a single session. A session can be considered as a period consisting of all interactions for the same information need. The category of a user's information need, previous queries, and recently viewed documents are all examples of short-term context. Such information is most directly related to the current information need of the user and thus can be expected to be most useful for improving the current search. In general, short-term context is most useful for improving search in current session, but may not be so helpful for search activities in a different session. The other kind of context is long term context, which refers to information such as a user's education level and general interest, accumulated user query history and past user click-through information; such information is generally stable for a long time and is often accumulated over time. Long-term contexts can be applicable to all sessions, but may not be as effective as the short-term context in improving search accuracy for a particular session [7].
- **Static and Dynamic Context:** Static context includes user profile, his interests and preferences whereas dynamic context includes user location, his current task, and vicinity to other people or objects.

2.1 Context Feature Dimensions for Context-aware web search

Based on the understanding of context typology, the context feature dimensions for context-aware focused web search system have been proposed (see Fig.5), which are discussed below:

- **User Profile**: It considers user background and his interaction with the system. It is further divided into three sub categories: *User Personal Information* (e.g. his interests, hobbies, behavior, etc.), *User Social Environment* (e.g. user interaction with other users, interests of a particular group of users), *User Tasks* (e.g. spontaneous activity, engaged activity, general goals, etc.)
- **Physical Environment**: Context related to physical environment is structured into two categories: location (absolute position, relative position, etc.), infrastructure (surrounding resources for computation, communication, task performance, etc.)
- **Time**: Situations and environments are generally characterized by a large degree of continuity over time, so that context history itself becomes an important feature for approximation of a given situation or environment. Time is implicitly captured in the history.
- **Queries**: Queries provided by the user also becomes an important feature in determining the contextual information. The analysis of nature of queries, keywords provided in the queries can provide clues in detecting the contextual information.

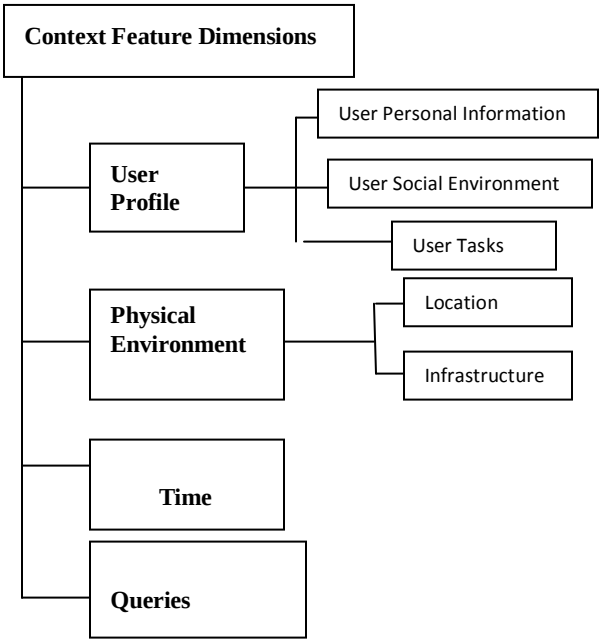

Fig. 5 Context Feature Dimensions for Focused Web Search System

3. Proposed Entity-Centric Context Model

In modeling the context, there is need to understand how to relate a situation to context. Since context awareness in the context-aware systems always relate to an entity, the contexts around entities can be characterized for a situation. An entity is a place, a subject, a device, an application, another context or a group of these. Modeling context around entities seems to be a natural way of building context-aware systems, because on the level of objects the concept of contexts and their relation to situations is in most cases well understood. Therefore, an *Entity-Centric model* is proposed in this work to understand and represent the context. This model is motivated with the idea that there is strong relationship between context and entities. The features of this model are discussed below:

- **Relation between Situation, Entity and Context:** When situation is modeled in the form of entities then it is easier to derive the context. Context can be described by understanding entities and their handling. Sensing on entity level carries additional information related to the domain knowledge of the entity.
- **Relationship between Entities:** There is inherent relationship between entities, especially when one entity is a part of another. Typically contexts represent the state of an entity or the relationship of an entity to other entities. It assumes that if an entity A is part of an entity B, then it is useful to know the contexts of A to determine the contexts of B. In reverse it is also assumed that the context of an entity can be estimated by knowing the contexts of all its sub parts.
- **Situation modeling through entities:** Since the entities are in typical situations, the context can be found out by monitoring typical situation related to the entities.

3.1 Representing Context in Context-Aware Focused web search

The representation of the context of a document may be presented with the help of a graph consisting of entities and their relationship. Generally, a document on the web contains a number of important keywords called entities related to each other. This relationship leads to the contextual information. For example, the web page shown in Fig. 6 consists of some important keywords that provide clue about the context of the page. The representation of these keywords and their relationship has been shown in Fig. 7.

…*Delhi*, the *capital of India* and major gateway to the country, is a fine blend of *ancient* and *modern India*. The antiquity of this great city dates back to the age of *Mahabharata* . *New Delhi* built by the imperial British is one of the most modern *metropolis city*. The historic city is full of *monuments, mosques, temples, gardens and museums*. *Best season to visit* is October to April. *Places of Interest*: Red fort, Jama Masjid, Jantar mantar,…….

Fig.6 Sample web page

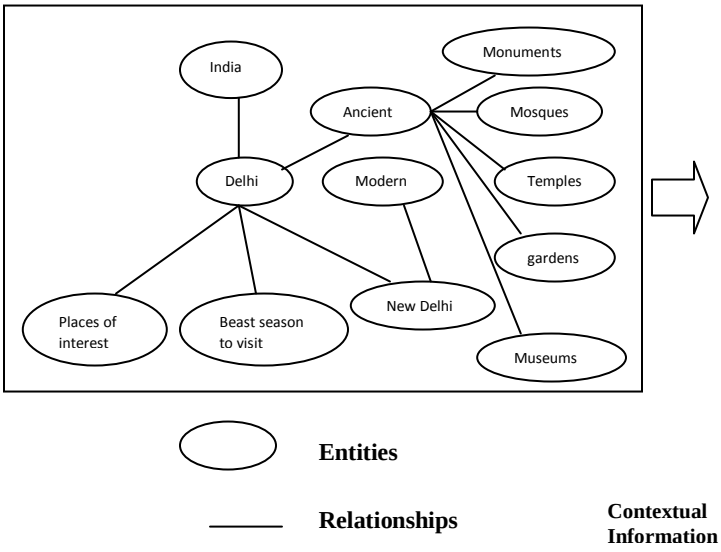

Fig.7 Context representation

4. Context Modeling Framework

A focused web search system aims to adapt the search system according to the user and environment contextual information so that it can recommend highly related web search results to the user. A general framework for modeling context in focused web search system is being proposed in this section. The framework for Context-aware web search system aims at the following features:

- Delivery of relevant search results based on the user's context thereby eliminating distractions and cognitive overhead related to the volume and level of information.
- Reduction in the user interaction with the system by using context as a filtering mechanism. This has the potential to increase the usability of web search systems by making them more responsive to the user needs.

The key components of the proposed framework are discussed below (see Fig.8):

- **Capturing the Context**

The purpose of this component is to identify and capture the contextual information from the user and his environment where he is working. The identification and capturing of context can be done in the following two strategies:

1. **Explicit Capturing**: In this strategy, the contextual information is provided by the user explicitly. In this method, each user may be asked to fill / answer some queries when the user gives his search query. The advantage of this method is that users interests are recognized directly and this explicit information given by the user helps in making *stereotypes* (user groups with common characteristics) [8] and *communities* (user groups with common interests) [9]. While explicit context information is more reliable, it is often not available because it requires extra effort from the user.
2. **Implicit Capturing:** In this strategy, the contextual information has to be implicitly judged from the user's behavior while he interacts with the system, his profile and environment [10,11].

- **Tagging Context along with Information**

The context information must also be tagged along with the information. For this, the following activities for the search system are required:

1. **Context Embedded Document:** The context of a web page document must be embedded in the document itself so that the search system while crawling on the web can compare the context of the topic to be searched with the context of the document being crawled. The context of the web page can be embedded either explicitly by the web page author or by automatic detection.
2. **Context-Sensitive Document Index:** The document index must also incorporate the context so that while searching the information from the index of the search system, the relevant document can be presented to the user.
3. **Context Enriched Query:** After capturing the context information, the context can be adapted: implicitly such that the user is not aware of this and he gets the highly relevant information after context information has been adapted to the system or explicitly such that the context derived is converted into the keywords form and augmented with the previous query of the user. After forming the new query with the context information, it is submitted again to the search system so that the user gets relevant results according to the context as shown in Fig..

- **Context Storage**

The context retrieved either implicitly or explicitly must be stored to maintain context history. Context history can be used to establish trends and predict context values. For example, the system may need the location history for a user in order to predict his future location. For these reasons the architecture of context aware web search system must support the storage of context.

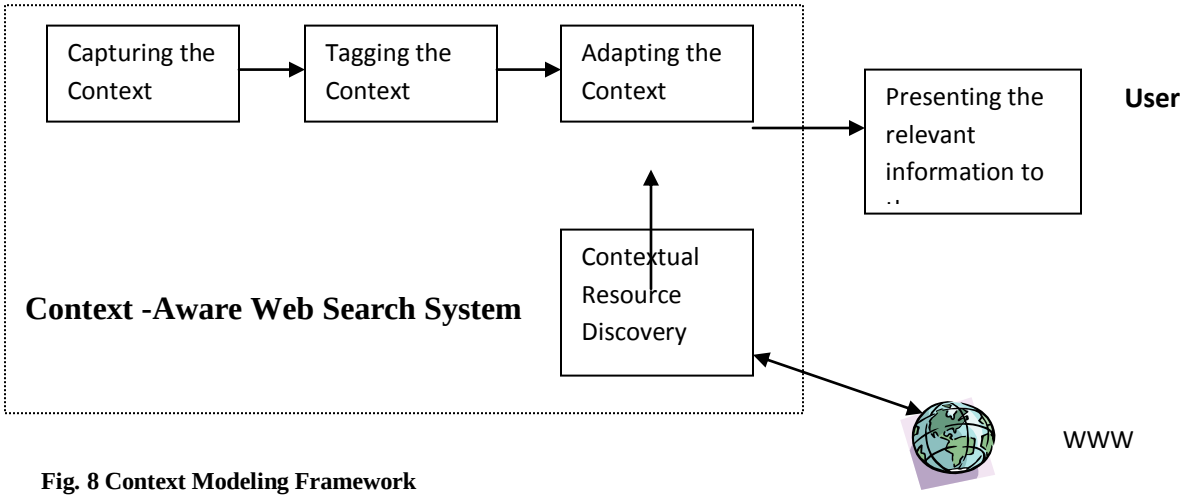

Fig. 8 Context Modeling Framework

- **Context Adaptation**

The web search system after capturing and learning the contextual information about the user and his environment must adapt itself to the user's actual needs presents highly relevant search results to him.

- **Contextual Resource Discovery**

Since the context aware web search system learns the user preferences and understands the area of his working, it may locate the information on the web, which is relevant to the user context and presents him with the recent updates on the required information. For example, a user is a research and working on the 'Focused Crawler'. The system learns the user context and may depute one agent, which searches the recent information on the web related to 'Focused Crawler'. The deputed agent collects recent information and sends back to the system periodically to be presented to the user.

Based on the framework presented, the context information flow for the proposed context-aware focused web search system is shown in Fig. 9.

The context-aware focused web search system attains the following benefits:

- The web search system will return only highly related web pages to the user.
- The system considers the user context including his preferences and interests.
- The system considers the environmental context like the user location.
- The system does not put cognitive load on the user in filtering the irrelevant search results.

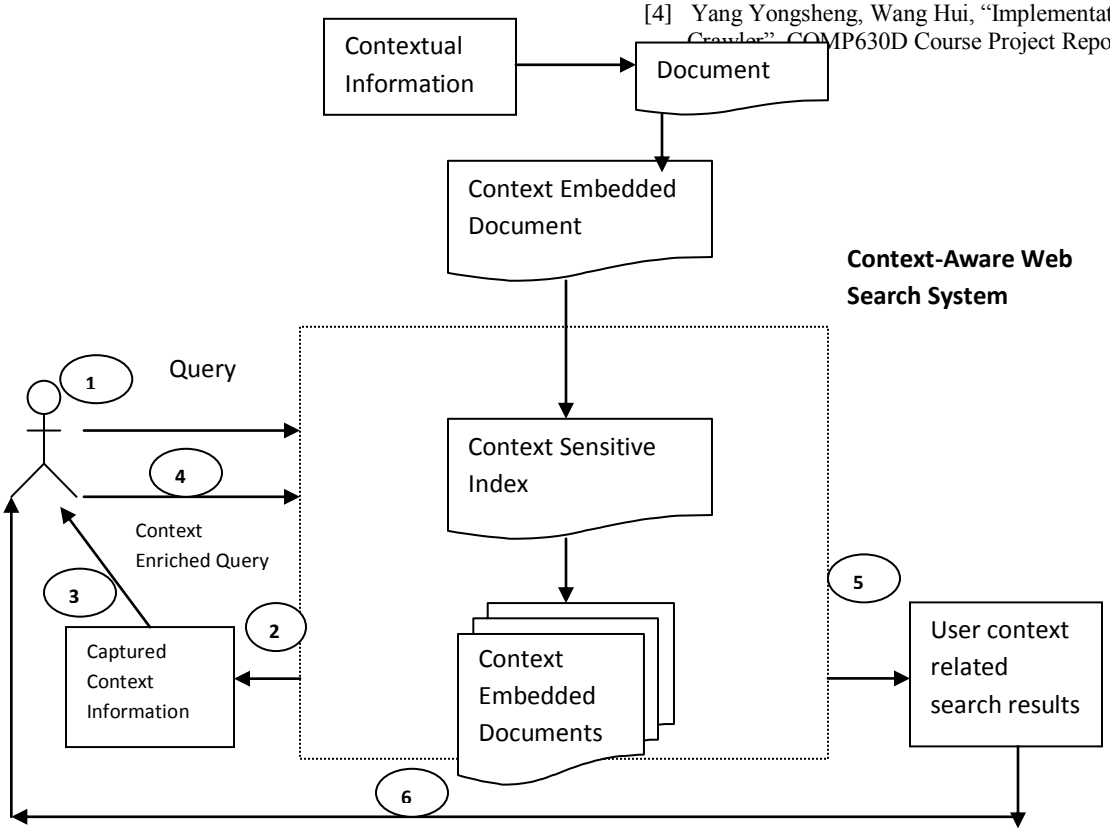

Fig. 9 Information Flow in Context Aware Focused Web Search System

Conclusions

In this paper, it has been emphasized that the context information is important for the web search systems so that they can retrieve only relevant web pages according to the user need. The context typology, context representation and an entity centric model have been discussed for the focused web search systems. A framework for context-aware web search system that incorporates the context model has also been proposed. The benefit of the context-aware search systems is that it reduces the cognitive load on the user in filtering out the irrelevant search results and return only highly related web pages to the user.

REFERENCES

[1] He D., Goker A., Harper D., "Combining evidence for automatic web session identification", Journal of Information Processing and Management, 2002

[2] S. Chakrabarti, M. Van Den Berg, B. Dom, "Focused Crawling: A New Approach to Topic specific web resource discovery", Proc. Of 8th International WWW conference, Toronto, Canada, May,1999

[3] Diligenti M., Coetzee F.M., Lawrence S., Giles C.L., Gori M., "Focused Crawling using context graphs", Proc. International Conference on Very Large Databases (VLDB '00), 2000,pp. 527-534

[4] Yang Yongsheng, Wang Hui, "Implementation of Focused Crawler", COMP630D Course Project Report

[5] James Allan et al, "Challenges in Information Retrieval and Language Modeling", Report of a Workshop held at the Center for Intelligent Information Retrieval,University of Massachusetts Amherst, September 2002

[6] Anind K. Dey, Gregory D. Abowd, "Towards a better understanding of context and context-awareness", Proceedings of the 1st international symposium on Handheld and Ubiquitous Computing, Karlsruhe, Germany, pp. 304 – 307, 1999

[7] Xuehua Shen, Bin Tan, C. Zhai, "Context-sensitive information retrieval using implicit feedback", Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005

[8] E. Rich, "Stereotypes and user modeling", User Models in Dialogue systems, A. Kobsa and W. Wahlster, eds. Springer-Verlag, pp. 35-51, 1989

[9] Paliouras G., Papatheodorou C., Karkaletis V., Spyropoulos C., Malavets V. "Learning user Communities for improving the services of Information providers", Lecture notes in Computer Science, 1513 : Springer-Verlag 367-384, 1998

[10]Naresh Chauhan, A.K. Sharma, "Smart agent based Recommendation system for Context driven Focused web search", Proceedings of 41st Annual convention of CSI (Computer society of India), Calcutta, Sep. 2006

[11] Naresh Chauhan, A.K. Sharma, "User modelling for adaptive learning in focused web search", Proceedings of National conference on Computational Techniques in Decision Making (NCCTDM), Bikaner, Oct. 6-8, 2006

[12] Naresh Chauhan, A.K. Sharma, "Analyzing user search trends on www", Proceedings of 2nd International Conference on Advanced Computing and Communication Technology, to be held at Panipat, Nov. 200727. Srivastava, A.N and Sahamp M 2009.Text mining,clustering and application ISBN:978-1-4200-5940-3 CRC Press

[13] SH. Zhash, P. Dmitriev, C. and Glies2009,Graph based crawler speed selection. www 2009, manrid, spain, ACM 978-1-60558-487-4/09/04

[14] Ahmadi-Abkenari, F. and Selamat A. July 2010. Application of click stream analysis in a Tailored Focused web Crawler. International Journal of Systemics and Informatics world network. Volume 10 PP:137-149.

[15].Castillo,Carlos(2004).EffectiveWebCrawling(http://chato.cl/research/crawling_thesis).(P.hd thesis). University of Chile . http://chato.cl/research/crawling_thesis). Retrieved 2010-08-03.