Information Retrieval in Web2.0

Ziran Zhang , Jianyu Tang
Information Managment Department, Central China Normal University,
Wuhan, 430079 China.imdec2003@163.com

Abstract. The development of Web2.0 not only update the network industry, but also vastly impact on the traditional retrieval methods of networked information and put forward more new demands. The paper analyses the development of Web2.0 and the new demand of networked information retrieval, describes and evaluates the current mode of networked information retrieval. At last, the author put forwards a new conceptual mode which based on JXTA and P2P of networked information retrieval.

1 Introduction

With the development of the web2.0 and its typical application such as Blog, wiki, RSS, Tag, SNS and so on since 2004, users have become the center of information production and usage and they have more point to point channels of information transmission. This development and change not only update the network industry, but also vastly impact on the traditional retrieval methods of networked information and put forward more new demands. Researching these demands and new retrieval progress would be helpful in providing better retrieval service for information users. This paper is intended to analyze these new requirements under the circumstance of Web2.0, discuss and evaluate new retrieval mode of networked information.

2 Development and Networked Information Retrieval Requirements of Web2.0

2.1 Traditional Retrieval of Networked Information

Traditional retrieval of networked information is mainly depended on such following methods: key words-based retrieval which is represented by search engine; subject catalogue-based retrieval which is represented by Yahoo's classification system; metadata-based retrieval which is represented by subject gateway, such as AHDS, EELS, MathGuide, etc [1]; database and relative data pattern and retrieval language based deep web retrieval; professional portal retrieval and so forth. However, most of them use professional data that are selected by website editors or created beforehand as the main data, so massive knowledge produced by public can't be processed; the retrieval results to different users' search are the same, so individualized service can't be realized; as lacking of semantic relation among information, so the search engine which is the most common ways of network search can get good retrieval results only to the objective and fact-based information, thereby the subjective and personal opinion-based retrieval can't be well done that high quality retrieval results couldn't be got or even there isn't any results [2], Such retrieval methods can't be satisfied obviously.

2.2 New Requirements of Networked Information Retrieval

In the environment of Web2.0, information takes on characteristics that differ from Web1.0 and consequently brings forward new requirements to information retrieval, providing possibility for the coming of new retrieval ways.

(1)The popularization and massive of information production

In the times of Web1.0, the production of networked information is centered on relative minority professional companies and website editors, whereas, a large number of users offer massive information content for network in the environment of Web2.0.However, as the constantly growing of Web2.0 users, the original networked information has assumed a tendency of blowout, which would undoubtedly take tremendous difficulties to users who want to get essential and valuable information from innumerable information.

(2)The microcontent and semantics of information structure

Microcontent comes from various data created by users, such as a network log, review, image, collected bookmark, preferable music list, things that want to do, places wishing to go, new friends and so forth. Web2.0 creates massive microcotent and consume as much microcontent everyday, therefore, how to help users manage, maintain, store, share, transfer microcontent become the key point that users whether could exploit information effectively or not [3]. Although there is lots of microcontent in Web1.0, such as album online, speak and response of forum and so on, which are relatively closed to outside that can not be organized and exploited again fundamentally, the applications in Web2.0 could reused these microcontent, which makes it's possible to use these microcontent freely in any place, so these microcontent could be aggregated, managed, shared and transferred, and be remixed to individualized and abound applications further [4]. Moreover, traditional web data in HTML has no semantic label and the transmitted semantics is identified by human, which may be greatly differs from the original meaning of information provider, this is also the fundamental reason that traditional methods of information retrieval usually can only use the simple retrieval ways of key words or Boolean among key words and combination of the two, so that the agreed point can not be gotten between information provider and information users. In the environment of Web2.0, Tag is a kind of semantic mark and there are other semantic labels, such as label of resource relationship, label that is endowed automatically by computers according to the usage condition of resources and so on. Although these semantic labels couldn't fully identified automatically by computers at present, semantic matching and clustering could be done as least [5], even users' retrieval habits and retrieval characters could be recorded on the basis of semantic labels provided and concerned by users.

(3)The good and bad information quality is more intermingled

The massive and dispersive information results in lacking of necessary control on information releasing, so the quality of information is more intermingled. For users of mastering different knowledge structure, how to discriminate, filter and utilize information becomes a big difficulty, and also increase the difficulty of identifying tools of information retrieval. In the meantime, the information semantic is more complicated because of the using of Tag: it's a unavoidable problem for retrieval application of the environment of Web2.0 that the degree of authenticity and accuracy of semantic labels supplied by information provider or information consumer whether could directly become the important evidence of information acquisition or not.

(4) The bidirectional of information spreading

As Web2.0 network being readable and writable, users' information feedback may be done at any time and information source can update information at any times, which is a real communication pattern of twin-channel and information transmission might be well discriminated namely the spread of multipoint to multipoint. Therefore, the traditional situation that users are passive to receive information and different users' requirements receive the same information retrieval results can be changed, and it's possible to further perfect the technology of push and pull.

In view of the above characteristics of information, networked retrieval in the environment of Web2.0 should be popular, social and individual: in the first place, there is enough information retrieval scope so that the retrieval to multiple information sources can be provided to meet the increasing of popular and substantial information in the environment of Web2.0; secondly, one-stop retrieval resembling the function of integrated search engine ought to be offered, so that users don't need to log on multiply websites to obtain similar information ,in the same time, the reuse and store of microcontent should be supported to make use of the technology advantage of Web2.0; thirdly, semantic relationship among information can be analyzed and offer certain service of information evaluation, filter and convergence for consumers according to the relationship, so that the difficulty of information judgment and selection faced by users can be reduced to a certain degree; fourthly, users are able to quickly acquire their interested information in time through search activities and drive their knowledge to update, the retrieval tools can record and evaluate users' search activities to keep accuracy of knowledge update and individual of retrieval results; fifthly, community is created for users' identify marks according to their search activities, high quality information source is pushed to users actively based on their interests. By using such retrieval, users could quickly find their preferable information, finishing their work or tasks or goals in life, sharing knowledge with friends or meet new friends by sharing, and enrich global information and knowledge bank by making use of "root wisdom"[2].

3 New Information Retrieval Patterns in the Environment of Web2.0

3.1 Extended Service of Search Engine

According to the above analysis, the future search should be social and the search engine ought to be more concerned with knowledge content produced by users, so that users' demands are met in the environment of Web2.0. Presently, lots of search engine extends the functions of Web2.0, such as forum search of Google, Blog search, knowing and space of Baidu, My Web2.0 of English Yahoo and so on. My Web2.0 is to create a "social" search engine, which searches the often used and trustful website content of professional group, so the weakness of massive search can be made up. When doing individual search, not only the special index of setting by yourself can be searched, the index set by your friends and workmates can also be searched. For example, a group of physicists may share individual set mutually, hence, when one searches, he could make sure the results he obtains come from the essential index set that's set and selected by others[6]; Zhongsou' s forum search, post bar and so on. Even the search engine Wikiseek which specially searches wiki comes forth [7]; there are still some companies, for instance, Qihoo directly positions its search developing direction in Web2.0 search. Although functions provided by search engine could in some extent help people get information in Web2.0, several problems still exit in it: (1) the search scope is relative narrow. The search extent limits to internal network, for example, forum in Google only search the forum information registered in Google; Baidu's knowing, the Knowledge Hall of Yahoo, Prefer to Ask Intellectual of Sina all only retrieve the knowledge content exchanged by net friends from individual database; (2) the retrieval channels are finite. Only a few attributes are offered to retrieve, for instance, Qiku's Blog retrieval provides only two accesses by Blog article and writer, lacking of the limitations from Blog website, comment on article and property category; (3) the retrieval results are short of semantic analysis. The retrieval results are still the same with traditional web search that lacks of semantic analysis so that users only get the sorting results by matching key words not the results according to different demands; (4) the retrieved information is short of essential filtration, which resulting in the unsatisfied retrieval results, even some garbage information is received.

3.2 Vertical Search Engine

Just as its name implies, vertical search engine is to search the information sources of one industry or a kind, simply speaking, it is to break up the big and comprehensive search engine to professional search engines. It's exactly because these applications of Web2.0 combining with the abundant, fast update and timely characteristics of information, causes the traditional search engine to be in difficulty, so it becomes reasonable to develop the vertical search engine [8]. At present, some comprehensive search engines such as Google, Baidu, Yahoo and so on, also enter into the area of vertical search engine, but there are more professional vertical search engines. The common are map search, music search, video search, news search, pal search, region search, science search, job search, classification information search, shopping information search, etc. In essence, many vertical search engines don't' belong to the category of Web2.0 search, as they are devoid of the core idea of Web2.0 which is users creating content. However, the vertical search engine brings lots of convenience to users' information searching undoubtedly, users use the vertical search engine to gain information closely linked with life and the channels of information spreading and communication are increased consequently. In the meantime, there are also some vertical search engines trying to combine with Web2.0 successfully, for instance, the Jobui[9] (http://www.jobui.com) utilizes the vertical search engine to settle information collection, each job hunter is helped to hunt for a suitable job by using Blog to introduce himself. Job hunters release their Blogs to be acquainted by recruit enterprises fully by network, which changes the previous situation that resumes are monotonous and dull, and so employment opportunities are increased; furthermore, the Blog conduct system is also opened, so job hunters can import the previous Blog in other website.

The biggest problem faced by the vertical search engine is how to deeply mining the professional information, analyze and record users' activities, and offer deep and professional information to appropriate users through the technology platform of Web2.0.

3.3 Networked Favorite

Networked favorite is a kind of free service of social bookmark, which is also called Cyber Digest such as del.icio.us and Yahoo collection+. Besides, there are still baidusoucang, drawer, collection bar, Shouker, Haowangjiao and so on offering networked favorite. The favorite have strong and also weak points, but following problems exist in them: (1) users' credit is lacked. Web2.0 stresses that users produce knowledge content which is the starting point of networked favorite, so that links, evaluation, abstract are all produced and shared by users. However, how to assess users with credit and so to confirm and recommend users by sort isn't considered by above tools. (2) The rationality of folksonomy can't be judged. Most of favorites create classification system according to tags given by users, which is regarded as the main channel of searching Shoucang of users, these classification systems are not to be compared with classification systems created by experts. (3)The retrieval functions are finite. Most of them only offer the search function through key words resulting from defining tag, retrieval functions according to attributes such as user, date, and region are not offered. (4)The functions of creating user community are limited. Although users of collecting the same material can be found, most of favorites don't offer service of building user community, which greatly reduces communication chances between users.

3.4 Personal Portal

With the constantly developing of Web2.0, the portal concept is mixed with search engine gradually, resulting in birth of personal portal tools, such as IG. IG (Internet/Information Gateway) is a kind of internet personal portal whose access is based on desktop, centers on personal user which is put forward by China search, realizing individual service which internet information is got, transferred and reactive timely. IG integrates multiple network service items, such as IE browse, QQ, MSN, search engine, information customization, individual information portal, website navigation, facilitating citizens service, BBS, entertainment service, etc. In IG, users may choose the search engine which is used for searching forum, Blog, MP3, etc, and subscribe their interested news, information, Blog and so on. In addition, MSN Live of Microsoft, Google pack of Google and space of Baidu belong to the software of desktop search. These search software offers much convenience to users' search and store information in the environment of Web2.0, but the problems mentioned above are not solved primarily, such as the problems of controlling of information quality, judgment of user rank, etc. Moreover, in IG's latest edition 2.1, the functions of search and subscription through key words only offer methods of post bar, news and forum, service of searching and subscribing Blog, wiki and other website's post bar, wiki isn't offered, so the service items are very finite. As a result, IG is just regarded as the traditional search engine of integrating Web2.0 and desktop tools.

4 The Conceptual Model of Networked Information Retrieval Based on JXTA and P2P

There are many retrieval patterns of Web2.0 at present, the key problem that users exploit information of Web2.0 has not been settled yet, such as user credit, information quality, extent of information sharing and so on. Hence, a conceptual model of networked information retrieval based on JXTA and P2P is put forward here.

4.1 P2P and JXTA

Traditional application patterns of network adopt the method of server to end computer, so users are passive to receive information from servers, but the new modes of P2P desalinate the boundary between service provider and user, so as to make each participant user become a supplier. P2P has the power of timely communication in network which is independent on device, and can look for and connect with the needed person as real time. Hence, from the aspect of human, the development point is not just referring to how to construct P2P network, but connecting people in network by P2P, so people could solve the exchange problems in the convenient network medium, this is what information exchange and exploit of Web2.0 need [10].

JXTA is a plan that Sun aims at constructing the general technology foundation of P2P, which defines a group of protocols, such as Peer Discovery Protocol, Peer Revolver Protocol, Peer Information Protocol, Peer Membership Protocol, Pipe Binding Protocol and Peer Endpoint Protocol. The system of JXTA is showed as Fig.1 [11].

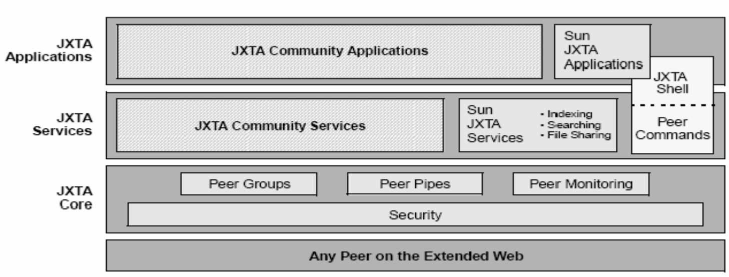

Fig. 1. P2P Software Architecture

4.3 The Conceptual Model of Networked Information Retrieval

The conceptual model of networked information retrieval based on JXTA is showed as Fig.2[12]. The JXTA Core and JXTA Services of the model can be fully referred to

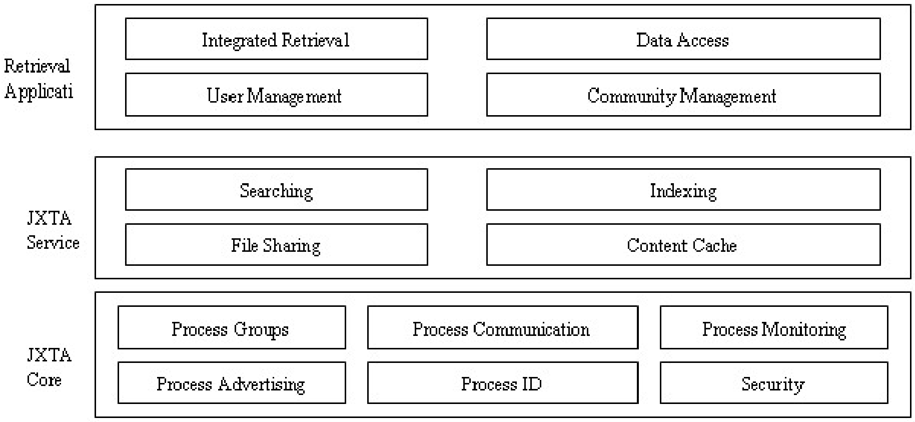

Fig. 2. The model of system structure

Retrieval Application mainly offers four kinds of functions such as integrated retrieval, data access, user management and community management. The integrated retrieval mainly provides retrieval request access, information filter, information retrieval and information sort. When a user or an information consumer submits a retrieval request, the request is submitted to system community and the set web search engine, such as Google, at the same time, the system searches corresponding results from process of community users or information providers who are related with the request, web search engine also returns retrieval results, and then the system integrates the results and feed back information consumers so that they can select the result that meet their needs. The selected files belong to the web or other users of the community who are information providers, the information consumer tags the files or directly chooses the tagged files and store then in his or her sharing process, so in the next retrieval process, information consumers are also information providers. Data access completes all the local data process and maintains a local cache space for community users' retrieval results. When there is same retrieval request, the local cache is searched firstly. User management is used for registration, login, allocating community, credit evaluation of users, so users can freely join in a community or are automatically classified to a community after the system analyzes their retrieval requests. Users' credit is evaluated according to the amount of files, the quality of tags and retrieval activity, the credit value is an important factor of sorting results when feeding back community retrieval results. Community management is used for managing sharing information of community in the system, such as users' past retrieval, community users and their credit and so on.

5 Conclusions

The paper analyzes these new demands in the environment of Web2.0, and evaluates and analyzes the tools and patterns of network information retrieval, and then a conceptual model of networked information retrieval based on JXTA and P2P is put forward. However, this is just the initial stage of the research, as the model is just a conceptual model. There is a lot of work to be done, such as how to extend scope of search, how to aggregate feedback information of different search circle, how to sort information effectively and so forth, which are all need to take great and continue efforts. At later time, we will do our best on the implementation of this model.

Reference

1. X. L. Zhang , "Semantics web and semantics-based networked information retrieval", *Journal of information*, 21(4),413-420 (2002).
2. Q. Zhang , "The search in times of web2.0"(January 10,2007). http://ysearchblog.cn/2006/12/web20.html.
3. "Microcontent in Web2.0" ,(November 18,2006). http://www.klogs.org/archives/2005/07/nieweb_20.html.
4. Incompleted, "Features of Web2.0"(November 18,2006). http://in.comengo.net/archives/feature-of-web2/.
5.W. Liu and. Q. Y. Ge , "From Web 2.0 to library2.0:service changed with users", *Modern information technology*, 9,8-12 (2006).
6. "Yahoo China lends Web2.0 activate to search mode"(December 1,2006). http://www.360doc.com/showWeb/0/0/159828.aspx.
7. "The catalog of search engine"(February 10, 2007). http://sskb.cn/sort/.
8. L. Liu , "Web2.0 of search engine. Journal of China Internet", 12, 32(2005).
9. Y. Liu., "Jobui: The new search of 'vertical search and Web2.0' ."(December 28,2006). http://net.chinabyte.com/chwssh/338/2232338.shtml.
10. W. T. Balke, "Supporting Information Retrieval in Peer-to-Peer Systems". *Springer-Verlag*, Berlin, Germany(2005).
11. G. Li., Project JXTA: A Technology Overview, 3(2002).
12.,M. Gnasa and Koblenz, "Congenial Web Search Conceptual Framework for Personalized", *Collaborative,Social Peer-to-Peer Retrieval*, MA, USA (2006).