Personalized RSS Search Service Using RSS Characteristics and User Context

Haesung Lee , Joonhee Kwon*

Abstract— **RSS is a one of the most important techniques in Web 2.0. Although there are a lot of RSS feeds available, finding which information are relevant to user isn't easy. The previous RSS search services have not taken into account RSS feed characteristics and user's contextual information. This seriously limits to offer users with useful information. This paper proposes a new personalized RSS search service using the RSS feed structure and user's context. The proposed method collects the data from RSS service site by categorizing RSS feed structure and then rank RSS channel using RSS tag characteristics and user's context. The system architecture and the search algorithms are described. We design a RSS feed crawler, RSS feed repository and a RSS feed search engine.**

Index Terms— **Context, Personalization, RSS, RSS feed search, Web 2.0.**

I. INTRODUCTION

Web2.0 is a new way to find, save, and share information generated on the web. The goal of Web2.0 is to facilitate forming network and sharing knowledge between users. Useful techniques that allow users easily to edit, store and publish contents on the web have been created. After that, multiple personal contents platform or personal media platform such as blog have developed. Using those platforms, a private individual publishes and shares his or her useful knowledge with others on the web very easily.

Web rapidly enlarges with explosively increased information. With the advent of Web2.0, information formed in UCC (User Created Contents) is created every day. The need of quickly obtaining and effectively sharing that information should be more concerned. In Web2.0, each user is provided with the different information what he or she interests individually and shares his or her own worth knowledge with others [1].

After appearance of the World Wide Web, initial search engines possessing only 1000 Web page indices, the method of acquiring information and ranking mechanisms have repeated a numerous change to provide more satisfactory services to users. In spite of dazzling technical development, the expansion of the web continually has demanded new and epoch-making techniques relevant the mechanism through which a search engine efficiently acquires information generated on the web and effectively provides each user with useful information [4].

The RSS (Really Simple Syndication) is regarded as next generation information delivery technique, which deliveries updated information to users simultaneously and frequently. Maximizing user's satisfaction in being provided with useful and reliable information is common goal of Web2.0 and RSS both [2].

Typically existing RSS search system offer user retrieval information as a result using mainly keyword based matching mechanism. But, the technical feature of RSS has is not suitable for algorithm of existing retrieval engines. Therefore specific prototype for RSS feed search systems is needed to offer each user not only the newest information but also interesting information.

One among the most time-consuming human activities in modern times is keeping up verified and useful data with a huge amount of continuously generated information. As a consequence of this overexposure, people's preference for some information has become one among the most valuable resource in Web2.0 to provide more personalized information to user. People's preference is very efficiently used to find and to acquire useful information among overexposure of information. As the scope of the Internet gets larger and larger, considering personalization becomes more and more important [1, 6].

In this paper, we propose a new RSS feed search system using user's context and feed characteristics of RSS. The proposed method collects the data from RSS service site by categorizing RSS tag data and then rank RSS channel using user's context and relevant tag value stored in the RSS feed repository. We mainly bring the RSS feed structure and user's context into focus to provide user with more reliable and personalized information. And to achieve this goal, it is needed to design the efficient web crawler specific to acquire RSS feeds not typical web document and RSS feed repository based RSS specification to improve the performance of RSS feed retrieval. Also, we introduce new ranking algorithm using various factors relevant characteristics of RSS and the context of each user.

Proposed RSS feed search system provide cool personalized, adaptive RSS feeds finding RSS service site for user based on her or his interests and other RSS feeds he or she read.

Manuscript received December 24, 2007. This work was supported by the Gyeonggi Regional Research Center.

* Joonhee Kwon is with the Department of Computer Science, Kyonggi University, San 94-6, Yiui-dong, Yeongtong-ku, Suwon-si, Kyonggi-do, Korea(corresponding author to provide e-mail: , kwonjh@kyonggi.ac.kr).

Haesung Lee is with the Department of Computer Science, Kyonggi University (email: seastar@kyonggi.ac.kr).

II. Related Works

The amount of the different information generated on the broad-scale web increases explosively. In Web2.0, it is very important that useful information for each consumer should be found, acquired and consumed effectively.

It is no longer the user that searches the information she is looking for, but it is the information she values that reaches directly its consumers [4]. RSS (Really simple syndication) is a very useful technique in Web 2.0 providing users a efficient new way to share content on their website with other users and makes it available to offer them various information without asking them to visit the site every time when new content is published. RSS can literally be used with just about any kind of web-based content. RSS fundamentally is a simple specification that uses XML and a format for web-based content in a standard way. RSS feeds are increasingly being used for other types of content. For example, you can get RSS feeds with weather forecasts, company news and financial information, package tracking and lots of others. RSS completely revolutionizes the paradigm according to which people collect information generated at number of information sources [3].

Although there are actually millions of feeds available, finding those that are appealing and relevant to a user isn't always easy. By rising necessity of RSS, consequently, search engines mainly dealing with RSS feed appear.

Fig. 1. BlogPulse

Although BlogPulse viewed at Fig. 1 is known primarily as a tool for tracking trends and hot topics in the blogosphere, it also has a good feed search engine, and depending on the numbers you believe, also has one of the largest indexes of feed-based content of any feed search service. BlogPulse's advanced search page provides phrase search, all the words or any of the words filters, and even allows a user to create her or his own free-form Boolean queries.

Fig. 2. Bloglines

Fig. 2 shows Bloglines. It is both a feed search tool and a feed reader/aggregator. A drop-down menu next to its search form allows a user to search all of the blogs it has indexed, only the blogs she or he subscribes to, the web, or add a feed to user's subscription.

They are all tiptoeing around RSS search, but none have yet to launch a full-blown RSS search service. While there are a number of smaller, specialized blog and feed search engines, their lack of resources and the problem of blog and feed spam mean their search results are often useless. So finding relevant feeds, at least for the time-being, often remains a hit-or-miss affair. Also, Those RSS feed search engines can provide effectively and quickly no useful results related each user's need for some information because they not mainly concern RSS feed structure and user's context.

Contexts are any information that can be used to characterize the situation of an entity [7]. An entity is a person, place, or object that is considered relevant to interaction between a user and an application. When humans talk with humans, they are able to use implicit situational information, or context, to improve more communicational functionalities [5].

Unfortunately, this ability to convey ideas does not transfer well to humans interacting with computers. By improving the application's access to context, it is available to increase the richness of communication in human-computer interaction and make it possible to produce more useful computational services [6]. There are various human factor related context such as knowledge of habits and emotional state of each user. To provide each user or searcher more interesting and useful information, number of application or solution has been developed using many various contexts such as user's preference [7, 8].

In Web 2.0, the concept of personalization is very important. As the scope of the Internet gets larger and larger, the need

for personalization to bring it within our scope becomes more and more important. But, any online examples where personalization was really well integrated with the user experience currently are very few. The best example we can come up with was Amazon's personalization engine and that is just scratching the surface of what's possible [5].

We had to reach out to a real-world analogy to try to understand the concept of personalization better: A store owner that knows you as a friend knows your personal likes and dislikes and always seems to find just the right thing for you. In fact, all you have to do is say what you're looking for in a few words, and based on how well he knows you, the store owner runs in the back and never fails to bring back exactly the right product, in the right size and the right color. The scenario describes where our proposed search engine wants to go.

Google provides a personalized service shown in Fig. 3, the gadget maker, which allows users to make their personalized web page by themselves reflecting their preference.

Fig. 3. Gadget Maker

Another example providing personalized service on the web is Pandora's music genome project offering users music streaming service based on their preference for music. This provides very useful service that allows users not to take the time of searching music related their preferences. With this service, user listens music related their preferences continually. The use of those contexts in user-centred service makes it possible to provide more useful and personalized information satisfying each user.

III. Personalized RSS feed Search System Using Characteristic of RSS and User's Context

To provide more relevant information to the user in RSS search service, it is important to incorporate the feed characteristics and the user's contextual information into the search process. However, the previous search services have not taken into account them. This seriously limits to offer users with useful information in RSS search service.

To solve these problems, we consider the following. Firstly, we use the RSS tag structure in search query. It enables users to limit search in a specific feed tag, not searching full RSS document. Secondly, we consider both the query term frequency and the update frequency in RSS channel. It enables users to get more useful RSS channel by considering multiple RSS channel features. Thirdly, we use the user's context in search service. To provide a personalized search service, it is needed to understand the user's interest or preference. The context has a large influence on the interest and intent of one particular user.

Considering those described above, we introduce personalized RSS feed search system in which two factors are mainly considered, characteristic of the RSS and user's context such as user's preferences. It is a new RSS feed search system that uses user preferences to match search results to their interests.

Fig. 4 shows the architecture of the personalized RSS feed search system proposed in this paper. The architecture is composed of RSS feed crawler, RSS feed repository, and RSS feed search engine.

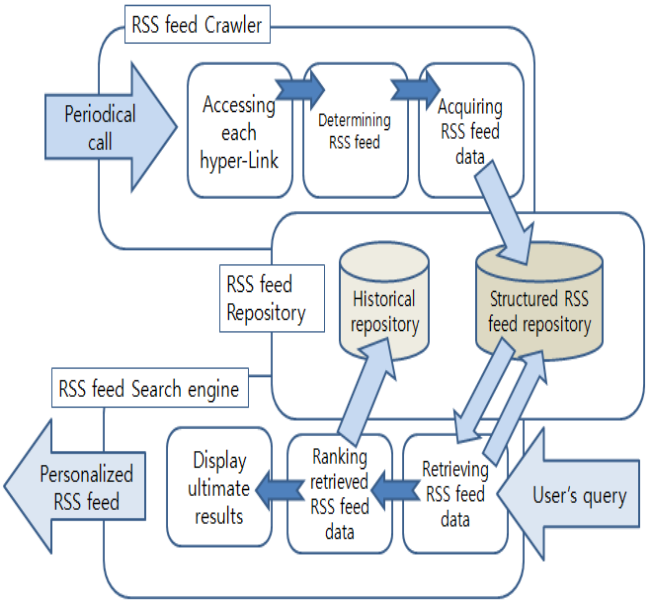

Fig. 4. System Architecture

RSS feed crawler visits RSS feed sites and reads their channels and then stores them in RSS feed repository. RSS feed repository stores RSS feeds composed of structurally in RSS feed database. RSS feed search engine receives user's search query and then returns RSS feed channel results from RSS feed repository to the user.

RSS feed crawler

Called periodically, RSS feed crawlers wander about on the web through the link to find RSS feed link. When accessing corresponded URL address, RSS feed crawlers judge whether the URL present RSS feed address or not. It is very important to acquire valid RSS feed address promptly. Our

each RSS feed crawler include feed information cash with which RSS crawlers directly check whether acquired RSS feed URL is duplicated to elevate the whole system's performance. If acquired RSS feed URL is verified as new RSS feed URL, RSS feed crawlers get the RSS feed URL to structured RSS feed database. Otherwise, RSS feed crawlers extract new contents from the RSS feeds.

Fig. 5 shows the operation flowchart of RSS feed crawler. By periodical calls which generate the crawler instance, the RSS feed crawler accesses site along hyperlink in the site and then downs web page to determine whether this page is RSS channel or not. If accessed page is RSS channel, RSS feed crawler checks duplicated accesses. RSS feed crawler acquired RSS feeds, and then insert it into structured RSS feed database.

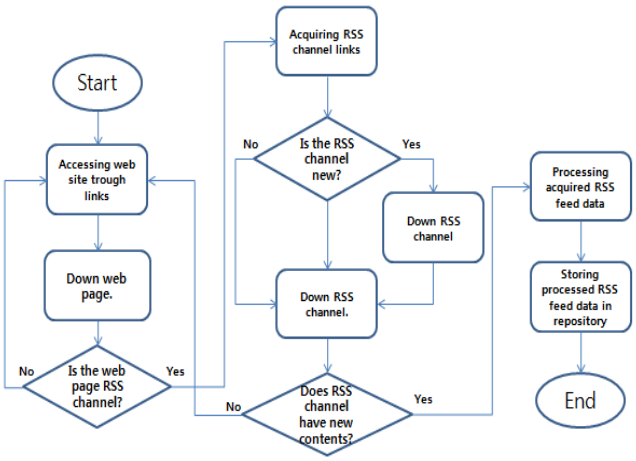

Fig. 5. Operation Flowchart of RSS Feed Crawler

RSS feed crawlers use depth-first search algorithm (DFS), which is used typically by many web crawlers [10]. Proposed RSS feed crawlers input links acquired from web pages to queue and draw each links out orderly to judge whether the link is RSS channel link or not.

RSS feed repository

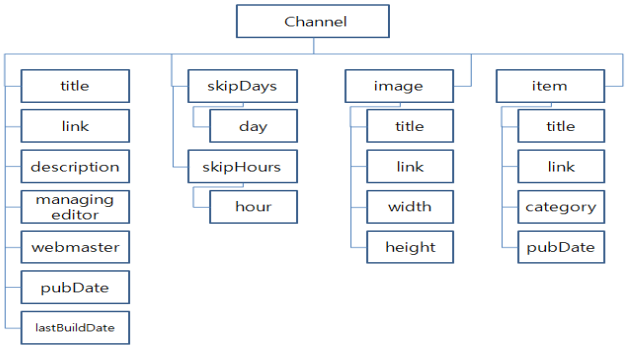

Fig 6. RSS Format's Structure

Fig. 6 shows a RSS format with number of tags. Those tags are composed structurally and expressed in the tree as viewed at Fig. 6. The use of those structural tags not makes it very simple to express the web contents but also makes the RSS feed data to access effectively.

Table 1 illustrates our RSS feed data schema. The RSS feed data are structured by sub tags of item and stored in structured RSS feed database. Structured RSS feed enables users to limit search in a specific feed tag, not searching full RSS document. This method can provide more useful RSS channel to users. For example, it allows users to limit results to content published within a particular date range, and sort results by date.

Table. 1. Sub tag of Item

Tag name	Description
title	The title of content
pubDate	The publishing date of content
author	The author of content
link	The link URL of content
description	Contents of updated feed

RSS feed search engine

In our proposed system, RSS feed search engine takes key role. The goal of proposed RSS feed search engine is to offer to users more useful information. We consider RSS feed characteristics and user's contextual information. The proposed search method is comprised of four main tasks.

First, a user enters a query based on structured RSS feed structure. Then, RSS feed search engine accesses RSS feed repository and look up stored contents based on user's query term frequency. It is commonly used method in traditional search engines.

```
m ← the number of published contents within sample period
total ← 0
temp[sample period]

for i ← 1 to sample period do
  if ( existing published contents at i )
  d[i] ← the publishing date of contents at i
 else
  d[i] ← 0

for j ← 1 to sample period do
  temp[j] ← d[j+1] − d[j]
  total ← total + temp[j]

return total / m − 1
```

Fig. 7. Update Frequency Algorithm

Second, we consider the update frequency in RSS channel. The reason of considering the update frequency of RSS channel is to conclude whether those RSS channels publish useful contents continually or not. We use the pubDate tag of each item of RSS feed to calculate update frequency. Fig. 7 illustrates the algorithm to calculate update frequency of the RSS channel.

Third, we conclude which content includes user's preference. By using user's context, it is possible to provide information with user's preference. Our RSS feed search engine takes the step in providing personalized search results based on user's preferences enabling searchers to specify what interests them. For it, proposed search engine organize user's historical repository. This repository consists of predefined user's profile, past visited site links or RSS channels. By using the historical repository, we can assume information that the user prefers. An example of user's preference is the RSS channels registered in the user's RSS reader.

Finally, our method ranks RSS channel using three tasks. Our ranking score is computed by (1). The RS_{uki} denotes the ranking score with respect to RSS channel i and term T_k of user u. It is possible to compute RS_{uki} with adjustable parameters α, β and γ $(0 \leq \alpha, \beta, \gamma \leq 1)$:

$$RS_{uki} = \alpha \frac{tf_{ik}}{tf_{max}} + \beta \frac{uf_i}{uf_{max}} + \gamma \frac{pd_u}{pd_{max}} \quad , \text{where } 0 \leq \alpha+\beta+\gamma \leq 1$$

tf_{ik} = frequency of term T_k in RSS channel D_i
tf_{max} = maximum frequency of term
uf_i = update frequency of RSS channel D_i
uf_{max} = maximum update frequency of RSS channel
pd_u = preference degree of user u
pd_{max} = maximum preference degree

(1)

After ranking RSS channels, RSS search engine provides result set of RSS channels in descending order by the computed ranking score.

Ⅳ. Conclusion

Web 2.0 is a new way to find, save, and share information on the Web. The goal of Web 2.0 is to facilitate forming network and sharing knowledge between users. In Web 2.0, RSS is a one of the most important techniques and a new way to provide a simple way for user to share contents on their website and make it available to users without asking them to visit the site every time when new contents are added. Although there are a lot of RSS feeds available, finding which information are appealing and relevant to user is not easy. Therefore, it is important to incorporate the user's contextual information and feed characteristics into the search process to get relevant information to user. However, the previous search system have not taken into account them when provide retrieved information as a result to user. This seriously limits to offer users with reliable and useful information.

This paper proposes a new RSS feed search system using user's context and feed characteristics of RSS. The proposed RSS feed search system efficiently collects the data from RSS service site by categorizing RSS feed structure and then rank RSS channel using user's context and relevant tag value. Using user's context, it does provide cool personalized, adaptive RSS feeds that automatically find RSS service site for user based on her or his interests and other RSS service site he or she read. That is, our proposed RSS feed search system uses personal preferences to deliver custom search results based on interests selected by users. Being different with the previous works, our proposed system acquires RSS feed on various web sites more efficiently using RSS feed structure and ranks information more reliably considering user's context such as preference, or predefined profile. Consequently, RSS feed search engines can learn from what a user do to help him or her find what he or she need. With integration of processes for each factor, we can efficiently rank the retrieval set of searching result by reliable RSS channel, the source of RSS feed. It makes it possible to provide users with more useful and personalized contents generated on reliable RSS channel.

Ⅴ. References

[1] Tim O'Reilly, What Is Web 2.0, 09/30/2005, http://www.oreilly.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html

[2] Ben Hammersley, "Content Syndication with RSS", O' Reilly & Associates, Inc, 2003. pp. 222, 2003.JC. J. Kaufman, Rocky Mountain Research Lab., Boulder, CO, private communication, May 1995.

[3] Chen Wu, and Elizabeth Chang, Aligning with the Web : an atom – based architecture for Web services discovery, Sevice Oriented Computing and Applications 2007, Vol. 1, Number 2

[4] Peter Briggs and Barry smyth, On the role of trust in collaborative Web search, Artificial Intelligence Review, 2006, Vol. 25, Number 1-2

[5] P.Bonhard and M.A Sasse, 'Knowing me, Knowing You' – Using profiles and social networking to improve recommender, BT Technoligy Journal 2006, Vol. 24, Number3

[6] Anind K. Dey, Understanding and Using Context . Personal and Ubiquitous Computing, Vol 5, Issue 1. 2001.

[7] Anind K.Dey, Gregory D.Abowd, "Towards a better understanding of context and context-awareness", Technical Report GIT-GVU-99-22, Georgia Institute of Technology, College of Cmputing, June 1999

[8] Albrecht Schmidt, Michael Beigl, Hans-W. Gellersen, "There is more to Context than Location", Proceedings Workshop on Interactive Applications of Mobile Computing, 1998.

[9] Paul Barford, Azer Bestavros, Adam Bradley and Mark Crovlla, Changes in Web client access patterns: Characteristics and caching implications, Word Wide Web, 1999, Vol. 2, Number 1-2

[10] Sergey Brin and Lawrence Page, The Anatomy of a Large-Scale Hypertextual Web Search Engine, Computer Science Department,. Stanford university, Stanford, CA 94305, USA. 1998 http://infolab.stanford.edu/pub/papers/google.pdf
W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.