THE POTENTIAL OF BOOKMARK BASED USER PROFILES

by

Asli Yazagan

Mathematics, Hacettepe University, 2006, Ankara, Turkiye

Information Science, University of Pittsburgh, 2010, Pittsburgh, PA, USA

Submitted to the Graduate Faculty of

School of Information Science in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2010

UNIVERSITY OF PITTSBURGH

SCHOOL OF INFORMATION SCIENCES

This thesis was presented

by

Asli Yazagan

It was defended on

04 29, 2010

and approved by

Stephen C. Hirtle, Professor, Graduate Program in Information Science and Technology

C. Michael Lewis, Graduate Program in Information Science and Technology

Michael B. Spring, Associate Professor, Graduate Program in Information Science and Technology

THE POTENTIAL OF BOOKMARK BASED USER PROFILES

Asli Yazagan, M.S.

University of Pittsburgh, 2010

Driven by the explosive growth of information available online, the World-Wide-Web is currently witnessing a trend towards personalized information access. As part of this trend, numerous personalized news services are emerging. The goal of this project is to develop a prototype algorithm for using bookmarks to develop a personal profile. Ultimately, we imagine this might be used to construct a personalized RSS reader for reading news online. A reader returns a large number of news stories. To increase user satisfaction it is useful to rank them to bring the most interesting to the fore. This ranking is done by implementing a personalized profile. One way to create such a profile might be to extract it from user's bookmarks. In this paper, we describe a process for learning user interest from bookmarks and present an evaluation of its effectiveness. The goal is to utilize a user profile based on bookmarks to personalize results by filtering and re-ranking the entries returned from a set of user defined feeds

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

The World Wide Web has had a profound impact on our everyday lives: we routinely rely on it as a ubiquitous source for timely information. In particular, the web's real-time and on-demand characteristics make it an ideal medium for news access anywhere and anytime. As a result, virtually every news organization now has a presence on the World Wide Web.

Driven by the explosive growth of information available online, the World-Wide-Web is currently witnessing a trend towards personalized information access. As part of this trend, numerous personalized news services are emerging. For example, Internet portals such as Google, Yahoo and Lycos offer personalized access to daily news stories from a large range of categories. These services are based on a static user profile; users fill out questionnaires in order to make use of news filtering capabilities. Personalization based on static user profiles is neither fine-grained enough to accurately reflect an individual user's interests, nor flexible enough to take a user's interest changes into account. In addition, it requires additional work from the user.

User profiles can also be constructed implicitly through agents that monitor user activity (Gauch, Speretta, Chandramouli, & Micarelli, 2007 p. 55). Time spent on a page, scrolling and mouse movement, eye tracking, downloading and bookmarking activities, browsing histories and web and search logs are all mechanism that can provide implicit feedback from user. Stoilova, Holloway, Markines, Maguitman, & Menczer (2005, pg. 67) claim that bookmarks contain explicit and implicit knowledge. K.Chan, (2000) and Stoilova, Holloway, Markines, Maguitman, and Menczer, (2005) propose techniques to define the level of user interest from user bookmark folders.

The next section provides information about web browser bookmarking and then discusses issues and motivation for extracting user interest from user bookmarks.

1.1 WEB BROWSER BOOKMARKS AND USER BOOKMARKING ACTIVITY

Bookmarks which are also known as “favorites”, are a tool available in all the major browsers. Browsers also offer various approaches to managing bookmarked URLs. Abrams, Baecker, & Chignell, (1998) define a bookmark as a recorded URL that may be saved in a browser for later use. Stoilova, Holloway, Markines, Maguitman, and Menczer, (2005) claim that bookmarks contain explicit and implicit knowledge. Explicit attributes are: URLs, titles, the hierarchical structure of the bookmark folder, bookmarked date and last accessed date, personalized title and description. Implicit knowledge can be developed by taking into consideration the way people organize or use these URLs. For example, folders and subfolders might be used to identify classification. Browser bookmarking facilities allow either a flat list or a hierarchical folder structure to store and organize URLs. According to Keller, Wolfe, Chen, Rabinowitz, & Mathe, (1997, p.1103), most browsers allow users to create a hierarchical organization of bookmarks.

Abrams, Baecker, & Chignell, (1998, pg.44) report about how users organize their bookmarks and the method they use. Twenty-six percent of users don’t organize their bookmarks. They leave them in the order they are created. When the number of bookmarks exceeds 35, 20% of users use folders. When the number of bookmarks exceeds 100, 20% of users use a hierarchy within folders. The number of folders follows a linear relationship with bookmarks. Users with the most bookmarks have more sophisticated methods to organize their bookmarks. For example, most users, who have more than 300 bookmarks, file bookmarked page when it is created and 50% of these users schedule a cleanup session to organize their bookmarks.

Abrams, Baecker, & Chignell, (1998, pg.47) report that bookmarks are used as “personal web information spaces” to help people remember and retrieve interesting web pages. According to the paper, bookmarks are created for (1) General usefulness, (2) Personal interest, (3) Quality of Web sites, (4) Frequency of use, (5) Potential future use. They also reports that bookmarks are used for: (1)

Identification: to make a spot to come back it later. (2) Collection: the ability to quickly retrieve sites. (3) Movement: remind where user was. (4) Episodes: Chronological list of episodes.

According to Abrams, Baecker, & Chignell, (1998, pg.46), bookmarks may be stored for short term or long term use. Abrams & Chignell reports that 30% of bookmarks have been visited in 3 months, 67% of them in 6 months and 96% of them in 1 year and 6% of them in 1 month. This means bookmarks are mostly used for archiving. This research also shows that users add bookmarks during periods of intense bookmarking separated by less active periods where few if any bookmarks are added.

1.2 ISSUES AND MOTIVATION

In this study bookmarks are used to generate a set of ‘interest clusters’. It is the goal of a larger research project to use. Generated interest clusters to rank or filter large amount information stores such as the entries from a set of RSS feeds. Given prior work on interest clusters this study looks to generate 5-11 interest clusters for a user. There are different categories of RSS feeds and user can get many feed entries related with one category in a small amount of time depending on the number of feeds user is subscribed. For example, if a user who is interested in politics did not check his reader for a day, he might probably get a large amount of information since political news is very dynamic. It would be nice to have ‘political interest cluster’ for user that gives the information about user’s political interest. Using the user’s ‘political interest cluster’ on the large amount of political news stories, user might be provided a better ranking of entries in the reader that saves time and increase user satisfaction. Bookmarks might be a good source to create the interest clusters because they can be considered as ‘personal interest archive’.

There are also some obstacles of using web browser bookmarks as a way to determine user interests. The ways to organize and describe bookmarks are limited on web browsers. Hierarchical folder organization is the only sophisticated way to organize them and text is the only descriptor for bookmarks. On the other hand, user might have bookmarks for short term use and never delete them. These might pose problems if we try to discern user interest from bookmarks.

1.2.1 Monolithic Structuring of URLs

Browser bookmarking facilities allow either a flat list or a hierarchical folder structure to store and organize URLs. Although superior to unstructured list, the widely-used hierarchical folder organization forces user to store information resources in disjoint, decomposable clusters, even when the information does not actually fit this pattern (Keller, Wolfe, Chen, Rabinowitz, & Mathe, 1997, p.1104). For example, a single web page describing restaurants in San Francisco may be useful in different circumstances and should be accessible by a number of different categories: restaurants, San Francisco, cooking, Web Page design, entertainment etc. (Keller, Wolfe, Chen, Rabinowitz, & Mathe, 1997, p.1104). Some browsers allow users insert a single bookmark into multiple folders by creating copies. Hierarchical folder organization makes it difficult to see connections among URLs (Keller, Wolfe, Chen, Rabinowitz, & Mathe, 1997, p.1104).

The simplistic structure of bookmarks can be a problem when extracting user interest. If a URL that can be mapped into multiple concepts is inserted in only one folder, then taking folders as interest clusters may not provide a good indication of user interest. Thus, even if users store bookmarks in hierarchical structure it may not be easy to extract user interest.

1.2.2 Lack of facilities to rank and conceptualize URLs according to utility

According to Abrams, Baecker, & Chignell, (1998, pg.44), most users agree that it is difficult to manage a large number of bookmarks. The presentation order of URLs in the bookmark folders in current web browsers is determined by initial user placement and users generally leave URLs in the order they are created (Abrams, Baecker, & Chignell, 1998, pg.42). On the other hand, bookmarks are represented by text obtained from the title of the page, but according to Abrams, Baecker, & Chignell, (1998, pg.47) it is not descriptive enough for users so that few users actually change the name of bookmarks because it is difficult to change the name to something more descriptive. As a result of the current state of web browser bookmarking facilities, user must work hard to continuously rearrange the folder content and order them to keep useful URLs and use them efficiently. This reality makes extracting user interest from bookmarks more challenging.

1.2.3 Noise in Bookmarks

Bookmarks can be stored for short term or long term use. Short term users may want to bookmark sites as a reminder where he/she was. This kind of usage causes noise in bookmarks when extracting user interest if the reminder is not removed from the list. Abrams, Baecker and Chignell (1998, pg.43) report that users often bookmark things like a newspaper front page with headlines and pages with links to the articles. This kind of bookmarks can introduce noise since they include multiple contexts and it is not clearly defined which one the user is actually interested in.

1.3 RESEARCH OBJECTIVE

The goal of this project is to develop a prototype algorithm for using bookmarks to develop a personal profile. Ultimately, we imagine this might be used to construct a personalized RSS reader for reading news online. A reader returns a large number of news stories. To increase user satisfaction it is useful to rank them to bring the most interesting to the fore. This ranking is done by implementing a personalized profile. One way to create such a profile might be to extract it from user's bookmarks. In this paper, we describe a process for learning user interest from bookmarks and present an evaluation of its effectiveness. The ultimate goal, beyond the scope of this thesis, is to utilize a user profile based on bookmarks to personalize results by filtering and re-ranking the entries returned from a set of user defined feeds.

1.4 ORGANIZATION OF THESIS

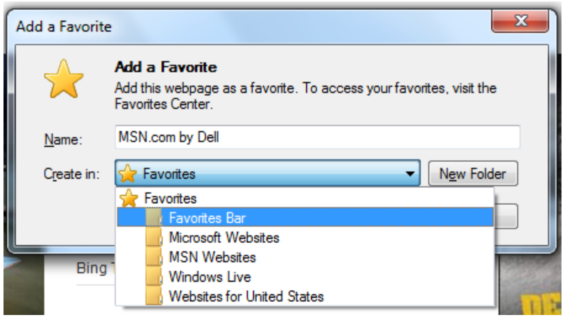
Chapter (-1-) presents an introduction to the study by discussing the context, the purpose of the study, motivation, issues and definitions. Chapter (-2-) reviews of the literature. Chapter (-3-) presents research questions and details of some of the preliminary work. Chapter (-4-) defines methodology of the study along the procedure of data analysis. Chapter (-5-) will provide the research findings and Chapter (-6-) presents implications of the research and recommendations for further research.

1.5 DEFINITION OF TERMS

1.5.1 Bookmarks

Bookmarks, which are also known as "favorites", are a tool available in the major browsers and provide an approach to managing URLs. When user wants to add a bookmark, they open a dialog window such as seen in Figure 1. The title of the page is selected as the bookmark name and a default folder is chosen.

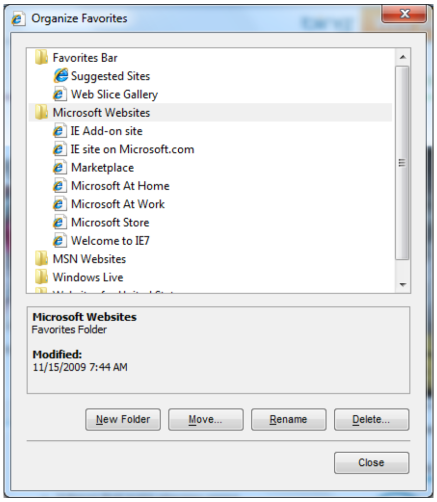
Users can customize name and folder at that time. It can be also organized later with a organization tool such as is shown in Figure 2.


Add a Favorite
Add a Favorite
Add this webpage as a favorite. To access your favorites, visit the Favorites Center.
New Folder
Favorites
Favorites Bar
Microsoft Websites
MSN Websites
Windows Live
Websites for United States
Bing

Name: MSN.com by Dell

Create in: Favorites

Figure 1. Add a favorite screenshot



Microsoft Websites
Favorites Folder

Modified:
11/15/2009 7:44 AM

Figure 2. Organize Favorites screenshot

Bookmarks can be exportable. A bookmark file has its own coding system. Each bookmark starts with <DT> tag and has anchor link (<A>) that has *href*, *add_date* and *last_modified* and *icon* and *icon_uri* attributes. This link is attached to the bookmark title.

```
<DT><A HREF="http://www.mozilla.com/en-US/firefox/help/" ADD_DATE="1264916415" LAST_MODIFIED="1269816889"
ICON_URI="http://www.mozilla.org/2005/made-up-favicon/1-1264916415518000"
ICON="data:image/png;base64”>Help and Tutorials</A>
```

Figure 3. An Example of Bookmark Code

1.5.2 RSS Feeds

An RSS (Really Simple Syndication) feed is an XML document that provides links to currently available content. At the top level, of an RSS document is the <rss> element, with a mandatory attribute called version, which specifies the version of RSS that the document conforms to. Subordinate to the <rss> element is a single <channel> element, which contains information about the channel (metadata) and its contents (Center, 2003). The XML excerpt below shows an example of a simple RSS 2.0 feed.

Title, link and description are the required channel elements. Title is generally the same as the title of the website. Link is the URL to the HTML website corresponding to the channel. Description is sentences describing the channel (Center, 2003).

```
<?xml version="1.0"?>
   <rss version="2.0">
      <channel>
         <title>Liftoff News</title>
         <link>http://liftoff.msfc.nasa.gov/</link>
         <description>Liftoff to Space Exploration.</description>
         <language>en-us</language>
         <pubDate>Tue, 10 Jun 2003 04:00:00 GMT</pubDate>
         <lastBuildDate>Tue, 10 Jun 2003 09:41:01 GMT</lastBuildDate>
         <docs>http://blogs.law.harvard.edu/tech/rss</docs>
         <generator>Weblog Editor 2.0</generator>
         <managingEditor>editor@example.com</managingEditor>
         <webMaster>webmaster@example.com</webmaster>
         <item>
            <title>The Engine That Does More</title>
            <link>http://liftoff.msfc.nasa.gov/news/2003/news-
VASIMR.asp</link>
            <description>Before man travels to Mars, NASA hopes to
design new engines that will let us fly through the Solar System more
quickly.  The proposed VASIMR engine would do that.</description>
            <pubDate>Tue, 27 May 2003 08:37:32 GMT</pubDate>
```

A <channel> may contain any number of <item>s. All elements of an item are optional, however at least one of title or description must be present (Center, 2003). The example above contains <title>, <description>, <link>, <pubdate> and <guid> elements. Title is the given title for the entry. Description is a short synopsis about the entry. Link is a URL link to the actual article. Pubdate indicates when the item was published. Guid is a string that uniquely identifies the item.

1.5.3 RSS Reader and RSS Aggregators

Really Simple Syndication (RSS) has a potential to changed how the web is explored by users. Essentially, RSS is a technology that lets web sites publish updates such as new blog posts or news articles. Interested users “subscribe” to the publication initiating a pull action where a user can read those updates which they have an interest in using a special *RSS feed reader* which can be either a desktop or Web-based application (Catona, 2009). Rather than forcing users to visit a Web site to check for new updates, the user can pull the updates to their reader.

The user subscribes to a feed by entering the feed's URL into the reader. Currently many aggregators allow the user to simply click an RSS icon in a browser to initiates the subscription process. The RSS reader checks the user's subscribed feeds regularly for new updates, downloads any updates that it finds, and provides a user interface to monitor and read the feeds (Wikipedia).

2.0 RELATED WORK

This chapter is divided into two parts:

- Section 2.1 reviews Profiles based on Web Data and Mining Bookmarks
- Section 2.2 reviews Personalization of News Services

2.1 DATA SOURCES FOR PERSONALIZATION

2.1.1 Profiles Based on Web Data

A user's profile is a collection of information that the system collects and maintains in order to improve the quality of information access. The goal of developing a user profile is to get the user more relevant information. User profiles may include demographic information, e.g., name, age, country, education level, etc, and may also represent the interests or preferences of either a group of users or a single person (Gauch, Speretta, Chandramouli, & Micarelli, 2007). User profiles can be constructed explicitly, through direct user intervention, or implicitly, through agents that monitor user activity (Gauch, Speretta, Chandramouli, & Micarelli, 2007, pg.55). Time spent on the page, scrolling and mouse movement, eye tracking, downloading and bookmarking activities, browsing histories and web and search logs can be used as implicit feedback from a user.

Different types of data can be used to create user profiles. WebMate (Chen & Sycara, 1998), PVA (Chien, Chen, & Yeali, 2004) and Syskill&Webert (Pazzani & Billsus, 1997) create user profiles by analyzing web pages visited by the user as they browse. The PVA system (Chien, Chen, & Yeali, 2004) uses web pages visited for more than 2 minutes to construct a user profile. OBIWAN (Trajkova & Gauch, 2004) uses web pages visited for more than 5 seconds and pages that are more than 1 KB and classifies them using tf-idf techniques. The highest 20 weighted words are used to represent the content of a webpage. Syskill & Webert (Pazzani & Billsus, 1997) is an agent designed to help a user with long-term information seeking goals, such as finding previously unvisited web sites on a particular technical topic. It uses a user-given topic name and the URL of an existing index page and user feedbacks to identify the interesting pages. An index page is simply a web page with many links to other pages on the topic. Syskill&Webert is not restricted to just pages that can be accessed from an index page. In particular, it has the ability to construct queries for search engines (consisting of a combination of highly informative words and words that occur frequently on interesting pages), and it can make suggestions about which pages returned from a search engine are of interest. The profile features are extracted from visited web pages. Do-I-Care (Starr, Ackerman, & Pazzani, 1996) is an agent which is designed to monitor user-specified web pages and notify the user of important changes to these pages. The user provides feedback on which changes are important from which the Do-I-Care Agent learns. The profile features are extracted from the difference between the current and an earlier version of a web page. Watson (Budzik & Hammond, 2000) gathers contextual information in the form of the text of the document the user is manipulating in order to proactively retrieve documents from a distributed information repository. They create a user profile by gathering terms based on rating documents using relevance feedback. Crabtree & Soltysiak (1998) uses received and sent emails and searched web pages to create a user profile. Speretta & Gauch (2005) creates user profile based on user search history by classifying the queries and snippets (titles and summaries) into concepts in a reference concept hierarchy.

These systems require a large amount of historical data and might be computationally expensive to mine. Further, when initially learning user interests, systems perform poorly until enough information has been collected to construct a reasonably complete user profiling. This is often referred to as the “cold-start problem.” It is contention of this study that bookmarks have valuable information in them which can be used to create user profile without gathering voluminous data over a longer period of time. Given the fact that most users have naturally accumulated bookmarks over time, the cold start problem can be overcome when sufficient bookmarks exist.

2.1.2 Mining Bookmarks

Several systems have been built that parse bookmark data and use it for personalization services.

PowerBookmarks (Li, Vu, Agrawal, Hara, & Takano, 1999) is a web information organization, sharing and management tool that parses metadata from bookmarked URL's and uses it to index and classify the URLs. Users can collect bookmarks into PowerBookmarks. The system monitors and utilizes user access pattern to provide personalization services such as automated URL bookmarking. If a bookmark is saved into PowerBookmarks, its URL is downloaded, its metadata such as keywords are parsed and it is indexed and classified. PowerBookmarks considers navigation history, association between URLs and existing bookmarks to recommend a URL to bookmark.

GiveAlink (Stoilova, Holloway, Markines, Maguitman, & Menczer, 2005) is an online bookmarking site. They determine similarity of bookmarks by mining the structure and attributes of bookmark files. They use Lin's measure (Stoilova, Holloway, Markines, Maguitman, & Menczer, 2005) to calculate similarity of bookmarks. The bookmarks are organized into tree structure for each user. Consider as an example a user u with bookmark x in folder F_x^u and bookmark y in folder F_y^u. Lowest common ancestor of x and y be folder $F_a{}^u{}_{(x,y)}$. Also let the size of any folder F, |F| be the number of bookmarks in that folder and all its subfolders. The size of the root folder is |U|. Then the similarity between x and y for user u is:

$$S_u(x, y) = \frac{2x \ \log|F^u_{a(x,y)}|/|U|}{\log|F^u_x|/|U| + \log|F^u_y|/|U|}$$

This function produces similarity values between 0 and 1. For example, if two bookmarks appears in the same folder, then their similarity is 1 because F_x=F_y=$F_{a(x,y)}$. Also, all other things being equal, the similarity between x and y is higher when F_y is a subfolder of $F_{x,}$ than when F_x and F_y are siblings (Stoilova, Holloway, Markines, Maguitman, & Menczer, 2005). Lin's measure works if bookmarks are organized in folders. If a user has a flat list bookmarks, according to user the Lin's measure, since all bookmarks are in the same folder all bookmarks must be similar, which is most likely not true. To avoid this problem, they decide to consider each bookmark in the flat list as if it were its own folder. Since Lin's measure calculates the similarity of bookmarks considering one user's bookmark set, in order to get a global similarity score for two bookmarks in the system, average score is calculated that are reported by each user. So that system can recommend URLs to the user.

Chirita, Olmedill, & Nejdl (2004) used bookmarks to improve the ranking of search results for a specific user. PROS takes bookmarks as an indicator for user preferences and defines a preference set based on bookmarks and pages surfed. They compare PROS algorithm with Page Rank and PPR (Personalized Page Rank). Page Rank algorithm considers a page that has many backlinks more important than others. PPR takes user selected pages as preference set. The result is that PROS returned the most relevant URL's among the three systems assessed.

Chan P. K. (2000, pg.42) describes a metric to estimate the level of user interest. The metric uses history, bookmarks, content of pages and access logs to estimate user interest. They proposed that if a user's bookmarks are organized into folders, then each folder can correspond to a cluster. Pages that are not bookmarked are assigned to the closest cluster measured by the average distance to each bookmarked page in the cluster. Ranking of pages are decided by using 1) number of pages visited by the user in a cluster during session and 2) how likely it is when a page is in a cluster is referenced.

Jung & Jo (2003) proposes a user-supported mechanism based on the sharing of knowledge with other users through the collaborative Web browsing, focusing specifically on the user's interests extracted for his or her bookmarks.

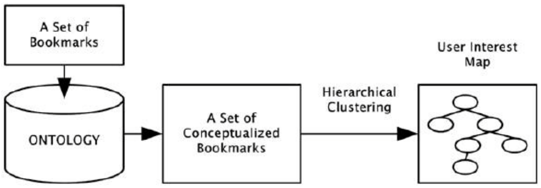


Figure 3. Interest extracting from bookmarks

Bookmarks are categorized on Bayesian networks by ontology. A bookmark set is replaced with category set for the user by referring to the ontology. The algorithm checks whether a category is connected to more than one category. If so, other categories are retrieved that are connected to it. This improves coverage of the user set as well as user preferences. Then it can be used to detect user preferences.

Hyoung-Rae & Chan (2008, p.153) extracts general and specific interest of a user that is called “User Interest Hierarchy (UIH)” based on user’s bookmark set. UIH represents user interest at different levels and these interests are learned from hierarchical clustering (DHC) to group terms that are extracted from bookmark content.

2.2 PERSONALIZATION OF AWERENESS SERVICES

2.2.1 SDIs, Ranking Platforms and News Services

Historically, "Selective Dissemination of Information" (SDI) has been used to refer to the matching of terms in a user-interest "profile" against document descriptors and selecting those documents with a specified degree of similarity to the terms of the user-interest profile (Connor, 1967). In SDI systems, modifications of the user profile are made manually by the user or the SDI staff. In most systems the users are encouraged to experiment with adjusting their interest profiles to produce the desired results (Connor, 1967). SDI is the foundation of current awareness services in that it is a personalized service, providing a specialized bibliography for each individual served, suited to his own research interests (Connor, 1967). The purposes of an SDI service are: (1) to provide a personalized current awareness service for the scientist keeping him informed of all research relevant to his interests, and (2) to conserve the time of the scientist by screening out irrelevant information, thus making the "information explosion" a manageable problem. The two traditional measures of the effectiveness of an Information Retrieval system, "recall" and "precision" are useful measures for the evaluation of an SDI system (Connor, 1967)

PROS (Chirita, Olmedill, & Nejdl, 2004) is a personalized ranking platform for web search. PROS takes bookmarks as an indicator of user preferences and defines a preference set based on bookmarks and page surfing, then extends this set to contain a set of hubs with high PageRank related to them. The PROS algorithm has been compared to Page Rank and PPR (Personalized Page Rank). The Page Rank algorithm considers a page that has many backlinks more important than others. PPR takes user selected pages as profile and creates hubs as PROS does. Experimental evidence found that PROS returned the most relevant URL's among three systems assessed.

2.2.2 Personalized News Services

Del Corso, Gulli, & Romani (2005, pg.98) proposed an algorithm to find the most authoritative news source and to identify the most interesting events in the different categories to which a news article belongs. The most important ranking criteria of their algorithm are freshness of news article and authoritativeness of the news agencies. They also claim that mutual reinforcement between new articles and news sources can be used for ranking.

iScore (Pon, Cárdenas, Buttler, & Critchlow, 2007a) aims to accurately predict interesting news articles for a single user. In iScore a variety of features are extracted from an article, ranging from topic relevancy to source reputation. The combination of multiple features yields higher quality results for identifying interesting articles, it also incorporates user-feedback.

Pon, Cárdenas, Buttler and Critchlow, (2007b) define "an interesting article" as an article that an arbitrary user finds interesting. The interesting classification task is to identify the most interesting articles from the entire pool of articles for different communities. The 43 RSS feeds considered for labeling are feeds of the form "most emailed", "most viewed", "most highly rated" and "top stories". For example, RSS feeds such as "Most reviewed Technology" is a good proxy of what the most interesting articles are for the technologist. They showed that better recommendation results can be achieved by tailoring the parameters to a specific user.

2.2.3 Adaptive News Readers

Krakatoa (Bharat, Kamba, & Albers, 1998) uses explicitly specified keywords as a part of the user profile. This provides a feature vector and each article is converted into a term / feature vector. Then all articles are ranked by their similarity score. WebClipping2 (Carreira, Crato, Goncalves & Jorge, 2004) used Bayesian classifier to calculate the probability that a specific article will be of interest to the user. The system uses implicit feedback which takes account of characteristics such as reading time. NewsDude (Billsus & Pazzani, 2000) creates short term and long term user models based on rated stories. Chan, Sun, & Lim (2001) have designed Categorizer to classify news articles from a news channel and it allows the user to create and maintain personalized categories. The user defines his/her personalized category by providing a name for category and a set of keywords associated with it. These keywords are known as the category profile for the newly created category.

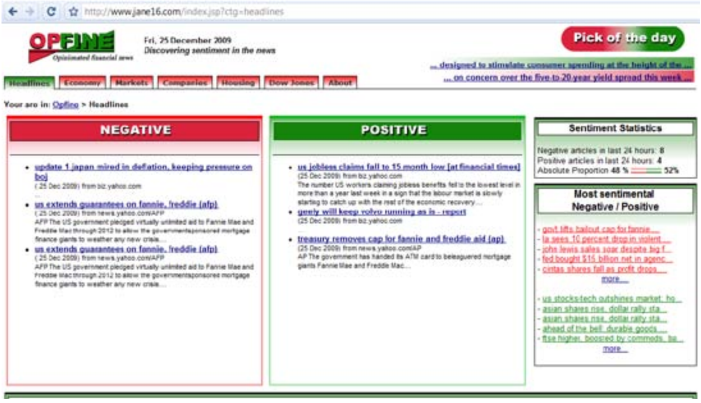
Table 1 Adaptive News Readers

System Name	Profile Construction Method
Krakatoa	Explicitly specified keywords
WebClipping2	User specifies their interests from available subjects with a value ranging from 0 to 100.
NewsDude	User model is created based on rated stories
Categorizor	User defines his/her personalized category by providing a category name and a set of keywords associated with it

2.2.4 RSS Aggregators

RSS allows people to subscribe to content and further it has allowed web developers to mash up news coverage in new and exciting ways (Catona, 2009). Sites like Regator, Techmeme and Opfine use RSS feeds to enhance the way readers find and consume the news.

Opfine (http://www.opfine.com) is a 100% automatic service collects the financial news stories from around the Internet. It reads the publicly available RSS feeds throughout the internet and analyzes their polarity (negative or positive). It displays them according to sentiment polarities - negative (red) and positive (green). Opfine uses the bespoke sentiment analysis algorithm, specifically designed and adjusted for this site to be financial terminology aware. Accuracy of Opfine is around 97%.



Fri, 25 December 2009
Discovering sentiment in the news

... designed to stimulate consumer spending at the height of the ...
... on concern over the five to 20 year yield spread this week ...

Your are in: Opfine > Headlines

NEGATIVE

- update 1 japan mired in deflation, keeping pressure on boj
  (25 Dec 2009) from biz.yahoo.com
- us extends guarantees on fannie, freddie (afp)
  (25 Dec 2009) from news.yahoo.com/AFP
  AFP The US government pledged virtually unlimited aid to Fannie Mae and Freddie Mac through 2012 to allow the governmentsponsored mortgage finance giants to weather any new crisis ...
- us extends guarantees on fannie, freddie (afp)
  (25 Dec 2009) from news.yahoo.com/AFP
  AFP The US government pledged virtually unlimited aid to Fannie Mae and Freddie Mac through 2012 to allow the governmentsponsored mortgage finance giants to weather any new crisis ...

POSITIVE

- us jobless claims fall to 15 month low [at financial times]
  (25 Dec 2009) from biz.yahoo.com
  The number US workers claiming jobless benefits fell to the lowest level in more than a year last week in a sign that the labour market is slowly starting to catch up with the rest of the economic recovery ...
- geely will keep volvo running as is - report
  (25 Dec 2009) from biz.yahoo.com
- treasury removes cap for fannie and freddie aid (ap)
  (25 Dec 2009) from news.yahoo.com/AP
  AP The government has handed its ATM card to beleaguered mortgage giants Fannie Mae and Freddie Mac ...

Sentiment Statistics

Negative articles in last 24 hours: 8
Positive articles in last 24 hours: 4
Absolute Proportion 48 % 52%

Most sentimental Negative / Positive

- govt lifts bailout cap for fannie ...
- la sees 10 percent drop in violent ...
- john lewis sales soar despite big f ...
- fed bought $15 billion net in agenc ...
- cintas shares fall as profit drops ...

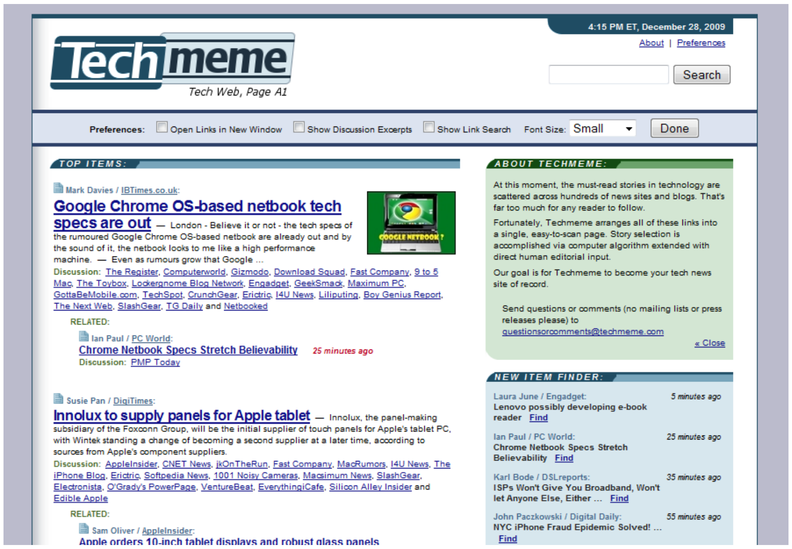
more ...

- us stocks-tech outshines market; ho ...
- asian shares rise, dollar rally sta ...
- asian shares rise, dollar rally sta ...
- ahead of the bell: durable goods ...
- ftse higher, boosted by commods, ba ...

more ...

Figure 4. Opfine Screenshot

Techmeme (http://www.techmeme.com) arranges stories in technology that are scattered across hundreds of news sites and blogs into a single, easy-to-scan page. Story selection is accomplished via computer algorithm extended with direct human editorial input.



4:15 PM ET, December 28, 2009

Preferences: □ Open Links in New Window □ Show Discussion Excerpts □ Show Link Search Font Size: Small

TOP ITEMS:

Mark Davies / IBTimes.co.uk:
Google Chrome OS-based netbook tech specs are out — London - Believe it or not - the tech specs of the rumoured Google Chrome OS-based netbook are already out and by the sound of it, the netbook looks to me like a high performance machine. — Even as rumours grow that Google ...
Discussion: The Register, Computerworld, Gizmodo, Download Squad, Fast Company, 9 to 5 Mac, The Toybox, Lockergnome Blog Network, Engadget, GeekSmack, Maximum PC, GottaBeMobile.com, TechSpot, CrunchGear, Erictric, I4U News, Liliputing, Boy Genius Report, The Next Web, SlashGear, TG Daily and Netbooked

RELATED:
Ian Paul / PC World:
Chrome Netbook Specs Stretch Believability 25 minutes ago
Discussion: PMP Today

Susie Pan / DigiTimes:
Innolux to supply panels for Apple tablet — Innolux, the panel-making subsidiary of the Foxconn Group, will be the initial supplier of touch panels for Apple's tablet PC, with Wintek standing a change of becoming a second supplier at a later time, according to sources from Apple's component suppliers.
Discussion: AppleInsider, CNET News, jkOnTheRun, Fast Company, MacRumors, I4U News, The iPhone Blog, Erictric, Softpedia News, 1001 Noisy Cameras, Macsimum News, SlashGear, Electronista, O'Grady's PowerPage, VentureBeat, EverythingiCafe, Silicon Alley Insider and Edible Apple

RELATED:
Sam Oliver / AppleInsider:
Apple orders 10-inch tablet displays and robust glass panels

ABOUT TECHMEME:

At this moment, the must-read stories in technology are scattered across hundreds of news sites and blogs. That's far too much for any reader to follow.

Fortunately, Techmeme arranges all of these links into a single, easy-to-scan page. Story selection is accomplished via computer algorithm extended with direct human editorial input.

Our goal is for Techmeme to become your tech news site of record.

Send questions or comments (no mailing lists or press releases please) to questionsorcomments@techmeme.com

NEW ITEM FINDER:

Laura June / Engadget: 5 minutes ago
Lenovo possibly developing e-book reader Find

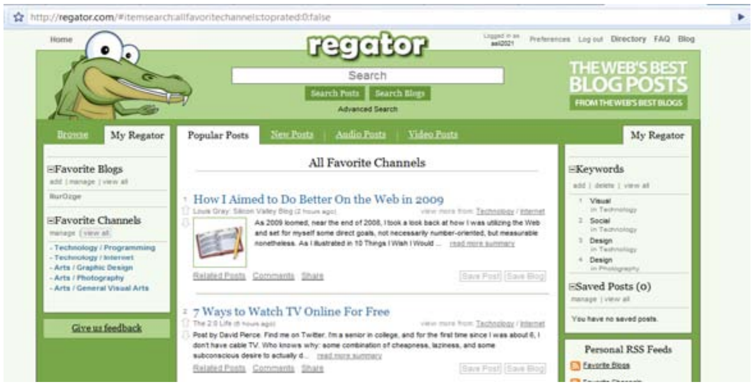
Ian Paul / PC World: 25 minutes ago
Chrome Netbook Specs Stretch Believability Find

Karl Bode / DSLreports: 35 minutes ago
ISPs Won't Give You Broadband, Won't let Anyone Else, Either ... Find

John Paczkowski / Digital Daily: 55 minutes ago
NYC iPhone Fraud Epidemic Solved! ... Find

Figure 5. Techmeme Screenshot

Regator (www.regator.com) is a website that gathers the world's best blog posts from Blogosphere and organizes them. Regator can also be used as an RSS reader. Users can upload their own selection of blogs and can search posts about any keyword. Regator can sort articles based on its popularity in the community or recentness.



All Favorite Channels

1 How I Aimed to Do Better On the Web in 2009

Louis Gray: Silicon Valley Blog (2 hours ago) view more from: Technology / Internet

As 2009 loomed, near the end of 2008, I took a look back at how I was utilizing the Web and set for myself some direct goals, not necessarily number-oriented, but measurable nonetheless. As I illustrated in 10 Things I Wish I Would ... read more summary

Related Posts Comments Share Save Post Save Blog

2 7 Ways to Watch TV Online For Free

The 2.0 Life (8 hours ago) view more from: Technology / Internet

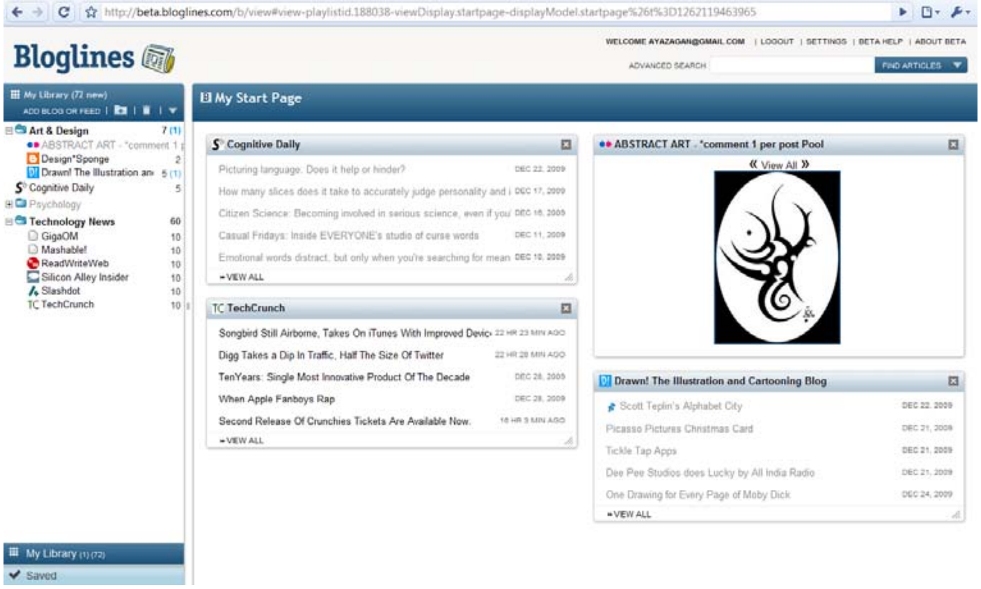
Post by David Pierce. Find me on Twitter. I'm a senior in college, and for the first time since I was about 6, I don't have cable TV. Who knows why: some combination of cheapness, laziness, and some subconscious desire to actually d... read more summary

Related Posts Comments Share Save Post Save Blog

You have no saved posts.

Figure 6. Regator Screenshot

Bloglines (www.bloglines.com) is an online service for searching, subscribing, creating and sharing news feeds and blogs. Users can create categories and make their own personalized page dragging and dropping feeds on the main page.


http://beta.bloglines.com/b/view#view-playlistid.188038-viewDisplay.startpage-displayModel.startpage%26t%3D1262119463965
Bloglines
WELCOME AYAZAGAN@GMAIL.COM | LOGOUT | SETTINGS | BETA HELP | ABOUT BETA
ADVANCED SEARCH
FIND ARTICLES
My Library (72 new)
ADD BLOG OR FEED
My Start Page
My Library (1) (72)
Saved

Art & Design	7 (1)
ABSTRACT ART - "comment 1 p	
Design*Sponge	2
Drawn! The Illustration an	5 (1)
Cognitive Daily	5
Psychology	
Technology News	60
GigaOM	10
Mashable!	10
ReadWriteWeb	10
Silicon Alley Insider	10
Slashdot	10
TechCrunch	10

Cognitive Daily

Picturing language: Does it help or hinder?	DEC 22, 2009
How many slices does it take to accurately judge personality and i	DEC 17, 2009
Citizen Science: Becoming involved in serious science, even if you'	DEC 16, 2009
Casual Fridays: Inside EVERYONE's studio of curse words	DEC 11, 2009
Emotional words distract, but only when you're searching for mean	DEC 10, 2009

» VIEW ALL

TechCrunch

Songbird Still Airborne, Takes On iTunes With Improved Devic	22 HR 23 MIN AGO
Digg Takes a Dip In Traffic, Half The Size Of Twitter	22 HR 28 MIN AGO
TenYears: Single Most Innovative Product Of The Decade	DEC 28, 2009
When Apple Fanboys Rap	DEC 28, 2009
Second Release Of Crunchies Tickets Are Available Now.	18 HR 3 MIN AGO

» VIEW ALL

ABSTRACT ART - "comment 1 per post Pool

« View All »

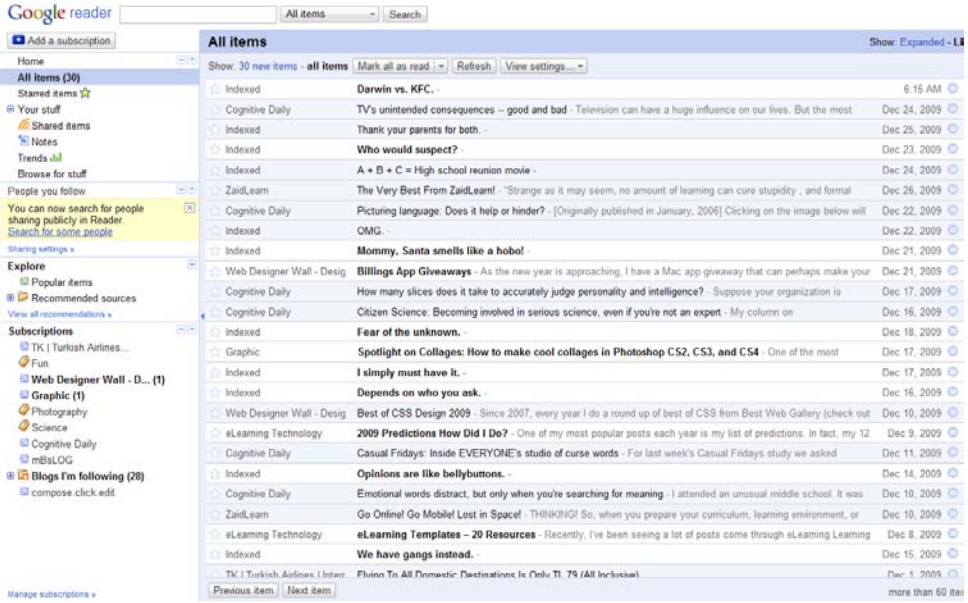
Drawn! The Illustration and Cartooning Blog

Scott Teplin's Alphabet City	DEC 22, 2009
Picasso Pictures Christmas Card	DEC 21, 2009
Tickle Tap Apps	DEC 21, 2009
Dee Pee Studios does Lucky by All India Radio	DEC 21, 2009
One Drawing for Every Page of Moby Dick	DEC 24, 2009

» VIEW ALL

Figure 7. Bloglines Screenshot

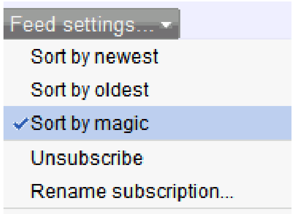
Google Reader (reader.google.com) allows a user to read and manage RSS Feeds. User can have folders and related feeds underneath them as Bloglines. Google Reader returns articles in chronological order by default.



Indexed	**Darwin vs. KFC.** -	6:15 AM
Cognitive Daily	TV's unintended consequences -- good and bad - Television can have a huge influence on our lives. But the most	Dec 24, 2009
Indexed	Thank your parents for both. -	Dec 25, 2009
Indexed	**Who would suspect?** -	Dec 23, 2009
Indexed	A + B + C = High school reunion movie -	Dec 24, 2009
ZaidLearn	The Very Best From ZaidLearn! - "Strange as it may seem, no amount of learning can cure stupidity , and formal	Dec 26, 2009
Cognitive Daily	Picturing language: Does it help or hinder? - [Originally published in January, 2006] Clicking on the image below will	Dec 22, 2009
Indexed	OMG. -	Dec 22, 2009
Indexed	**Mommy, Santa smells like a hobo!** -	Dec 21, 2009
Web Designer Wall - Desig	**Billings App Giveaways** - As the new year is approaching, I have a Mac app giveaway that can perhaps make your	Dec 21, 2009
Cognitive Daily	How many slices does it take to accurately judge personality and intelligence? - Suppose your organization is	Dec 17, 2009
Cognitive Daily	Citizen Science: Becoming involved in serious science, even if you're not an expert - My column on	Dec 16, 2009
Indexed	**Fear of the unknown.** -	Dec 18, 2009
Graphic	**Spotlight on Collages: How to make cool collages in Photoshop CS2, CS3, and CS4** - One of the most	Dec 17, 2009
Indexed	**I simply must have it.** -	Dec 17, 2009
Indexed	**Depends on who you ask.** -	Dec 16, 2009
Web Designer Wall - Desig	Best of CSS Design 2009 - Since 2007, every year I do a round up of best of CSS from Best Web Gallery (check out	Dec 10, 2009
eLearning Technology	**2009 Predictions How Did I Do?** - One of my most popular posts each year is my list of predictions. In fact, my 12	Dec 9, 2009
Cognitive Daily	Casual Fridays: Inside EVERYONE's studio of curse words - For last week's Casual Fridays study we asked	Dec 11, 2009
Indexed	**Opinions are like bellybuttons.** -	Dec 14, 2009
Cognitive Daily	Emotional words distract, but only when you're searching for meaning - I attended an unusual middle school. It was	Dec 10, 2009
ZaidLearn	Go Online! Go Mobile! Lost in Space! - THINKING! So, when you prepare your curriculum, learning environment, or	Dec 10, 2009
eLearning Technology	**eLearning Templates – 20 Resources** - Recently, I've been seeing a lot of posts come through eLearning Learning	Dec 8, 2009
Indexed	**We have gangs instead.** -	Dec 15, 2009
TK | Turkish Airlines | Interr	**Flying To All Domestic Destinations Is Only TL 79 (All Inclusive)**	Dec 1, 2009

Figure 8. Google Reader Screenshot

Google Reader has feed entry sort option called "magic" that re-orders items in the feed based on user's personal usage, and overall activity in Reader instead of the default chronological order.



2.3 SUMMARY

This chapter looked at data sources for personalization and following the overview of profiles on web data, bookmarks are introduced as the data that is used for personalization purpose. Also SDIs, personalized ranking platforms and news services, standard and adaptive RSS readers which are the latest personalized information retrieval technologies are reviewed.

Current adaptive systems require either a large amount of history data which is computationally expensive to mine or user-defined profiles which can be a burden on the user. Further, when initially learning user interests, systems perform poorly until enough information has been collected for user profiling. It is contention of this study that bookmarks have valuable information in them which can be used to create user profile without the need to wait a long period of time until enough implicit data is collected.

3.0 RESEARCH STATEMENT

This study seeks to determine whether a user's bookmarks can be used as a source of information that will reflect a user's interests. Ultimately, but beyond the scope of the current study, we are interested in whether integrating bookmark information into a user profile will improve personalized presentation order of RSS entries.

3.1 RESEARCH QUESTION

The following are important research questions that need to be addressed:

1) Can bookmark information be converted into profile made up of a set of Interest Clusters?
2) How to measure the quality of the created Interest Clusters?

3.2 PRELIMINARY ANALYSIS

As suggested in the literature, bookmarks saved by a user may be mined in variety ways. They may be organized or not, etc. This study reports preliminary investigations one whether a user's bookmarks can be used as a source of information that will reflect a user's interests. We conducted a number of informal reviews of bookmarks to better understand their structure and examine how a bookmark set might be converted into user profile.

3.2.1 How to Describe a Bookmark

A bookmark is some text and a URL that points to a website. When users bookmark a website, the default name of the bookmark is the title of the webpage. Most bookmark tools allow the user to modify this name. It is assumed by many that user modifications of a bookmark title make it more important than the default name. Seven sets of bookmarks were analyzed to make a preliminary determination of the number of modified titles. In the seven sets of bookmarks, there were a total of 858 bookmarks. The number of bookmarks in each of the seven sets was 30, 44, 76, 108, 229, 293.We found several potentially interesting things. Of the 858 bookmarks, 100 bookmarks titles were modified. Of those, 70 were shortened version of the original title. For example, the title “Books –Blogcritic” was modified to “BC Books” which does not bring any value. We suspect that shortened titles were deemed too long to be in the bookmark list conveniently. The following table show some examples of original titles of resources vs. user modified titles.

Table 2. Original Titles of Resources vs.. User Modified Titles

URL	TITLE	USER GIVEN TITLE
http://www.etiquettehell.com	Etiquette Hell — Your site has redeemed the web and my faith in humanity that there are some basic fundamental rules to life and personal relationships and that the poised will someday be victorious over the tacky heathens of the world	Etiquette Hell
http://www.adaa.org/	Anxiety Disorders Association of America ADAA Anxiety Disorders are real serious and treatable America ADAA Anxiety Disorders are real serious and treatable	ADAA homepage
http://blogcritics.org/books/	Books - Blogcritic	BC Books
http://health.msn.com/medications/default.aspx?vendor=google?match_type={ifsearch:search}{ifcontent:content}?pkw=%20medication	Medications | Prescription Drugs - MSN Health & Fitness	Drug Finder - MSN Health & Fitness
http://www.youtube.com/watch?v=5VoYHtAGLns	Peaceful Concentration Relaxation Meditation-Music-Part2	Concentration
http://www.indiaparenting.com/pregnancy/data/preg32_00.shtml	Pregnancy	Pregnancy week-by-week
http://levin.senate.gov/newsroom/supporting/2008/Detainees.121108.pdf	http://levin.senate.gov/newsroom/supporting/2008/Detainees.121108.pdf	csas detainee report

The 30 modified titles that were not shortened bookmarks were analyzed to understand reasons of modification. Of 30 bookmarks, 14 of bookmarks’ titles were the URL of the resources – which is often the case if the resource is not an HTML page e.g. it is a PDF which was the case for 11 of the 14 pages or an image file which was the case for 3 of the 14 pages. 4 of the 30 bookmarks were in another language so that user modified it.

Table 3. Detailed Analysis on Modified Titles

Number of Bookmarks	Number of modified bookmark titles	Number of modified titles that are the shortened version of the original title or user added terms those are not descriptive	Other modifications (modifications that are different from url original title or user adds term that are descriptive)	NON HTML PAGES of other modifications
30	-	-	-	-
44	2	-	2	1 (pdf)
76	16	9	7	5 (pdf)
78	1	1	-	-
108	19	13	6	1 (pdf)
229	18	16	2	1 (pdf)
293	44	31	13 (3 in other language)	6 (3 pdf, 3 image)

Remaining 16 bookmarks of 30 non-shortened modified bookmarks sometimes had descriptive terms for the resources; sometimes user just shortened it using his words. Here are some examples:

Table 4. Valuable Modified Title Examples

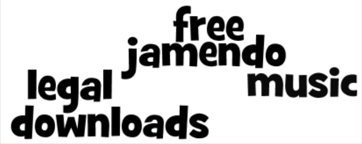
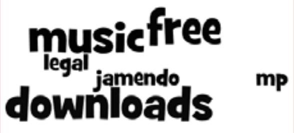
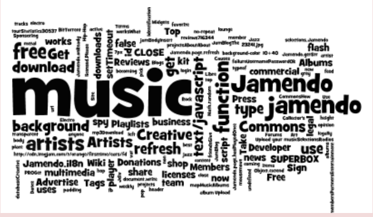
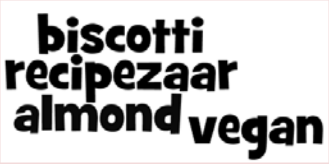
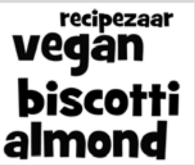
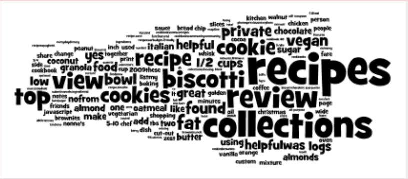
TITLE	USER GIVEN TITLE
Buttonator	Button Creator
Justices Hear Case on Right to Choose Defense Counsel - New York Times	nyt scalia twinkie defense
Judge of the Day: Diane Boswell - Above the Law - A Legal Tabloid - News, Gossip, and Colorful Commentary on Law Firms and the Legal Profession	cell phone contempt

Based on this analysis, we believe it will be true that modified bookmarks will be more representative of the topic of interest when a URL is replaced with a title, but are not more descriptive when the title is merely shortened.

Since modified titles would not bring that much valuable information about the URL, its body text and only the original URL title are considered to extract representative terms for a URL. In order to decide the best and efficient way to extract representative terms for a URL, all body text, title text and combination of title text and H1 title text in the body are compared.

The Table 5 shows tag clouds of various combinations of information sources title text, title and h1text and all body text. Most of the time, title and h1text have the same terms. It might be argued combining h1 text brings nothing valuable, but actually even though they are very similar, it appears to promote more relevant terms. It is also possible that title and h1 text are exactly the same so that having h1 will not make any sense but also it will not harm. On the other hand, taking account all body text brings too much noise and does not describe a concept. Based on all this analysis, we believe that a combination of title and h1 text will be most representative of a URL.

Table 5. Tag Clouds of Various Combinations of Information Sources Titles

TITLE	TITLE + H1	ALL BODY TEXT
http://www.jamendo.com/en/		
		 Top works free Get download downloads setTimeout false CLOSE Reviews kit get flash Albums commercial music text/javascript function Jamendo Press type jamendo background spy Playlists business Commons legal Creative artists Artists refresh Take Developer use script news SUPERBOX Jamendo.i18n Wiki Donations shop Members Sign multimedia player share licenses now Free Advertise Tags team uses
http://www.recipezaar.com/32416		
		 kitchen walnut person sauce bread chip slices chicken private chocolate people peanut inch used italian helpful cookie vegan sugar share change yes together whisk coconut print recipe 1/2 cups recipes cookbook granola food cup 2009these low view bowl biscotti baking coffee top notes nofrom cookies great golden review minutes friends almond one oatmeal like found page javascript vegetarian christmas wide brownies make shopping nonna's 5-10 chef add tbs two fat collections dish cut-out zest butter oven using helpfulwas logs vanilla orange almonds custom mixture
http://www.gardeningpatch.com/herbs/growing-basil.aspx		

<table>
<tr>
<td>

growing
know
basil
need
grow

</td>
<td>

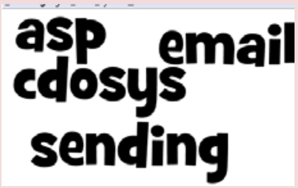
advice
know
grow
growing
need
basil

</td>
<td>

compost
gardening
outside
sowing
week
leaves
indoors
soil
dry
basil
pots
plants
help
stems
growing

</td>
</tr>
<tr>
<td colspan="3">http://www.w3schools.com/ASP/asp_send_email.asp</td>
</tr>
<tr>
<td>

asp
email
cdosys
sending

</td>
<td> </td>
<td>

report
ms
free
use
database
seo
access
data
affiliate
box
form
query
one
date
myform
button
internet
engine
wizard
source
add
filter
news
richard
tips
vba
sales
criteria
text
script
http://s7.addthis.com/js/addthis
txtstartdate
txtenddate
parameters
increase
make

</td>
</tr>
</table>

3.2.2 K-Means Clustering on Bookmarks

K –means clustering classifies a given data set by partitioning n vectors into k clusters. This is accomplished by assigning each vector to a set of randomly assigned points that serve as centroids for subsets of the data set. The location of the centroid is moved to the center of the assigned vector and then the closest vectors are again assigned to these points. This process continues until the sets don’t change and or the center points stop moving (Segaran, 2007, p. 33)

One goal of this research is to convert bookmarks into interest clusters. In order to create these clusters, K-means clustering is used to cluster user bookmarks. Using the following process:

1. For each bookmark URL, a representative vector will be created based on its text content as described in section 3.2.1

2. Using K-Means clustering algorithms, these vectors are clustered into k subsets with respect of their similarity.

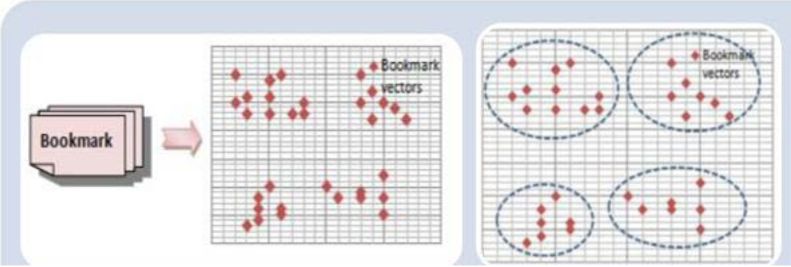


Figure 9. K-means Clustering on Bookmarks

In order to go further, it is necessary to decide how many clusters should be created with K-Means clustering algorithm so that they best define points of interest for user.

A bookmark set is clustered into k subsets where k is 5, 7, 9 or 11. In order to decide wheather a cluster set is a good estimation of user interest or not, user evaluation is needed. It is our hope that we will find particular levels of influenced by the number of bookmarks or the use of folders, or some other variable.

3.2.3 Quality Measures for Clusters

A cluster consists of some number of vectors where each vector represents a bookmark. All bookmarks in the cluster might be from the same folder, or some of them might be from different folders. It is assumed that as much as a cluster gets a good portion of bookmarks from the same bookmark folder, it represents a concept that is more likely an interest for user. One goal of this exploratory research is to determine if human judgement can be predicted based on objective measures of bookmark sets.

For quality measure of a cluster the following variables are defined:

CBM_{Fi} = Total count of bookmarks in the cluster that belongs to folder i.

BM_{Fi} = Total count of bookmarks in folder i

The proportion of CBM_{Fi}/ BM_{Fi} is called folder score for folder i and donated as $FS_{\mathbf{i}}$**.**

$$FS_i = \frac{CBM_{Fi}}{BM_{Fi}}$$

Folders whose FS >= 0.4 are called ‘dominated folders’ and donated as **DF**.

Folders whose FS < 0.4 are called ‘noise folders’ and donated as **NF**.

Cluster Quality Score (CQS) is calculated as below:

$$CQS = \frac{\sum_{i\ in\ DF} FS_i \ - \ \sum_{i\ in\ NF} FS_i}{Count\ of\ DF}$$

For example, a cluster has 9 bookmarks from folder A that has total 9 bookmarks and 2 bookmarks from folder B that has total 10 bookmarks. FS_A is 1 and FS_B is 0.2. In this case, A is a dominated folder (DF) and B is a noise folder (NF). Calculated CQS is 0.8 for this cluster. Since the cluster has all bookmarks from A, CQS is high and so high CQS indicates it is a good cluster. 2 bookmarks from B do not create a negative effect on CQS since they are from a NF.

For another example, a cluster has 4 bookmarks from folder A that has 5 bookmarks, 5 bookmarks from folder that has 27 bookmarks and 2 bookmarks from folder that has 13 bookmarks. FS_A is 0.8, FS_B is 0.18 and FS_C is 0.15. In this case, A is a DF and B & C are NF. Calculated CQS is 0.47. Since this cluster has bookmarks from different folders, its quality is not considered as good as previous example.

Average of the Cluster Quality Scores for a k-cluster set is defined as Cluster Set Quality Score and denoted as $CSQS_k$. This score is tentatively proposed as a measure of the best clustering set between 5, 7, 9, 11 cluster sets.

$$CSQS_k = \frac{\sum_{i=1}^{k} CQS_i}{Count\ of\ DF}$$

CSQS formula depends on the folder scores so it is important to define what a ‘*folder’* means in a bookmark set. A bookmark set might have linear folders without any subfolder, as example user bookmark set A is organized into 8 folders and these 8 folders do not have any subfolders. This kind of bookmark set is called bmset with 1-level folders. A bookmark set also might have folders in hierarchical structure with many subfolders. Each subfolder creates a level. This kind of bookmark set is called a bmset with k-level folders where k is the level of subdirectories. As an example, let say bookmark set B is organized into 5 folders. Of the 5 folders, 2 folders have 3 subfolders in them. Of 3 subfolders, 1 of them has 3 subfolders. This bookmark set is called a bmset with 3-level folders.

The CQS formula is design for bookmark sets with 2-level folders. If a bookmark set has more than 2 level folders, then it is converted into bookmark set with 2-level folders using the following rules:

- If a 1-level folder has only subfolders, each subfolder is considered as a ‘folder’.
- If a 1-level folder has one or more bookmarks and some subfolders, 1- level folder is considered as a ‘folder’ with all bookmarks that are not in the subfolders. Each subfolder is also considered as ‘folder’.
- If a folder’s subfolder has a subfolder, in other words, if a bookmark set has 3-level folders, then all 3-level folders’ bookmarks are combined and added to 2-level folder.

Let say folder A has a subfolder B and x number of bookmarks. B has 2 subfolders C and D and some number of bookmarks in C and D. Folder A is considered a ‘folder’ with x number of bookmarks and B is considered another ‘folder’ with all bookmarks of C and D.

In order to see how much different the CSQS score when 3-level folders are taken account or not, three bookmark sets that have 3-level folders were analyzed. Cluster sets are created and cluster scores are calculated taking account 2-level subfolders and also 3-level subfolders. Cluster Set Quality Scores are compared. The Figure 9 shows that there does not appear to be any significant effect of using a bookmark set with 2-level or 3-level folders on CSQS.

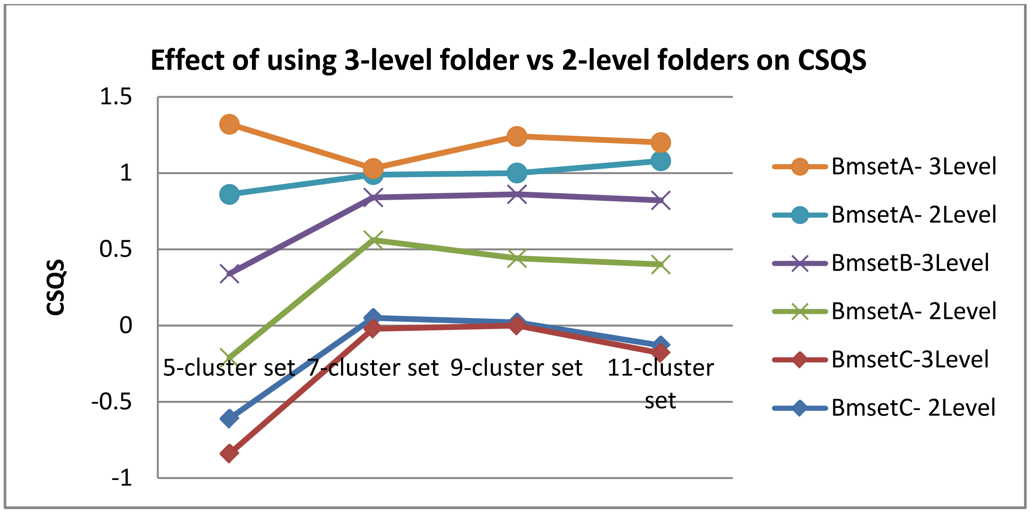


Figure 10. Effect of using 3-level folder vs 2-level folders on CSQS

A bookmark set that has 185 bookmarks and 17 folders has been classified into 5, 7, 9 and 11 clusters with K-means clustering algorithm. All bookmarks in this set are organized into folders, 32 of them are organized in sub folders. Table 6 shows the calculated cluster set scores for clustering level. According to the table, 9-Cluster set is expected to be the best cluster set for user interest. (Keep in mind, the goal of building and assessing a CSQS is to validate its use as a predictor of human judgment.)

Table 6. Calculated Cluster Set Scores for Clustering Level

Cluster set	**Cluster Set Quality Score (CSQS)**
5	0.11
7	-0.10
9	0.24
11	0.14

In order to see how well the best interest cluster is predicted using CSQS, users were asked to rate the clusters that were created with 5, 7, 9 and 11 clustering process. Tag clouds of clusters were created as visual representation of them. Clusters are displayed in random order to the user who is asked to rate clusters on scale of 1-7 (1 is less relevant, 7 is most relevant). Based on user relevancy ratings of clusters, average score is calculated for cluster set quality score for user. This score is called **cluster set relevancy score** and denoted as **CSRS**. Table 7 shows the User Relevancy scores for each clustering level. According to CSRS scores, 9-Cluster set is the best as it is expected.

Table 7. User Relevancy Scores for Each Clustering Level

Cluster set	**Cluster Set Relevancy Score (CSRS)**
5	4.2
7	4.0
9	4.7
11	3.3

Another bookmark set has 293 bookmarks in 75 folders. All bookmarks in this set are organized into folders, in other words there is not any bookmark that is not in a folder. The bookmarks were clustered into 5, 7, 9 and 11 clusters using K-means. Table 8 shows the system predicted scores for each clustering level. According to the table, 5-Cluster set is expected to be the best cluster set for user interest.

Table 8. System Predicted Scores for Each Clustering Level

Cluster set	**Cluster Set Quality Score (CSQS)**
5	0.81
7	0.62
9	0.62
11	0.65

User is asked to rate the randomly displayed cluster tag clouds that are created with 5, 7, 9 and 11 clustering process. Table 9 shows the user relevancy scores for each clustering level. According to CSRS scores, the 7-Cluster set, not the 5-Cluster set, is the best. This is not predicted, but do note that 5 and 7 cluster sets are closer to each other than the 9 and 11. While the results are not strongly encouraging, we think there is still some promise here.

Table 9. User Relevancy Scores for Each Clustering Level

Cluster set	**Cluster Set Relevancy Score (CSRS)**
5	4.6
7	4.85
9	4.0
11	4.2

4.0 METHODOLOGY

This research proposes to use bookmarks to create interest cluster sets and a method to measure the quality of interest clusters sets using bookmark hierarchal structure. Short of a controlled experimental study, we carried out further exploratory analysis of several more bookmark sets

4.1 DEVELOPMENT OF CLUSTER SETS BASED ON USER BOOKMARKS

Based on preliminary analysis, we continue to use a combination of bookmark title and h1 text on the body of the page is the most representative text for a bookmark as well as the most efficient way to describe it. For each bookmark, a representative vector was created based on its text content and vectors will be clustered using K-means clustering algorithm. This process has been achieved with a java program. The steps of this process are as follows:

1) All bookmark URLs, their titles and if exist their H1 texts are saved in MySQL database for a given username. If the bookmark is in a folder, folder name added to the record

2) All bookmark titles and H1 texts are retrieved and terms are extracted to create dictionary. All stop words and adjectives are removed

3) For each bookmark, its title and h1 text are combined removing all stop words, adjectives and digits as representative text. Bookmark vector is created based on the dictionary. If a dictionary term appears in bookmark representative text then frequency of the term is placed on the vector.

4) After all bookmark vectors are created, K-means clustering is performed using 5, 7, 9, 11 clusters. All clusters created at end of the each clustering process are called **k-cluster set** where k is 5, 7, 9, or 11.

4.2 EXPERIMENTAL DESIGN

In section 3.2.3, Cluster set quality score metric (CSQS) is explained.

$$CSQS_k = \frac{\sum_{i=1}^{k} CQS_i}{Count\ of\ DF}$$

The study examines the quality of this metric to predict the best cluster set for the user.

4.2.1 Predicting the Best Clustering Level for Users

The study examined user satisfaction with created interest clusters to determine whether CSQS score could return the same clustering level as the most qualified level for users. The participants were volunteers who have at least 70 bookmarks that were organized in folders. Users provided bookmarks for cluster construction and evaluated the created clusters. The system returned clusters in randomized order and they were displayed to the participants for relevancy rating. User were asked to rate the relevancy of the clusters on seven-point scale; where ‘1’ means not interesting at all, ‘7’ means extremely interesting. Before rating the relevancy, subjects were informed that the results would be displayed in a random order. The subjects in this experiment are considered experts on their interests. Their relevancy ratings for each cluster are considered perfect.

The relevancy ratings of each cluster were used to calculate cluster set relevancy score (CSRS). Average of the clusters’ scores that are belongs to the k-cluster set is defined as CSRS $_k$ for k-cluster set.

$$CSRS_k = \frac{\sum_{i=1}^{k} UserRatings_i}{k} \quad \text{where k} \ = \ 5, 7, 9, 11$$

Based on the CSRSs, cluster sets will be ranked for the perfect ordering.

4.2.1.1 Evaluation Procedure

To evaluate the performance of the proposed CSQS metric, the proportion of bookmark sets that have highest CSRS and CSQS for each clustering level will be compared.

In our experiment, we will examine whether the best clustering level based on CSQS score and CSRS score is the same or not.

4.2.1.2 Expected Result

It is expected CSQS and CSRS return the same clustering level as the level that is rated most highly by the participants.

4.2.1.3 Analysis Procedure

After user relevancy ratings, average user relevancy scores (CSRS) are calculated for each clustering level. Based on the calculated score, clustering levels are ranked from the highest score to lowest for each user as perfect ordering of user satisfaction indicating the best clustering level that creates best clusters defining user interest points. Number of users is counted for each clustering level that is obtained as the best clustering level according to CSRS rankings of each user. Clustering level with the best portion of number of user is decided to be the best clustering level.

At the same time, CSQS are calculated for each clustering level and ranked from the highest score to lowest for each user. Number of users is counted for each clustering level that is obtained as the best clustering level according to CSQS rankings of each user. Clustering level with the best portion of number of user is decided to be the best clustering level.

At the end of the calculations, obtained two clustering level are expected to be the same. This would indicate that CSQS can be used of deciding the best clustering level that provides the most likely the best cluster set would satisfy users.

5.0 RESULTS AND DISCUSSIONS

This chapter describes the result of the experiments.

5.1 PARTICIPANTS AND OVERALL PROCEDURE

The assessment was carried out between February and April 2010. The participants were individuals who have at least 70 bookmarks that were all organized in folders. Nine students from University of Pittsburgh applied to be participant. All participants provided their bookmarks for cluster construction and evaluated created clusters. Bookmarks are parsed and their titles and H1 texts are extracted and combined for representation of a bookmark set. Then, K-Means clustering is performed on it. It has been classified into 5, 7, 9, 11 cluster sets. 5, 7, 9, 11 are called clustering levels in the section. A total 32 clusters were created (5+7+9+11). They were converted into tag clouds. Each cluster tag cloud provides visual representation of potential interest clusters for user. These tag clouds are provided to the user in randomized order for relevancy rating. Users were asked to rate the relevancy of the clusters on seven-point scale; where ‘1’ means not interesting at all, ‘7’ means extremely interesting. Before rating the relevancy, subjects were informed that the results will be displayed in a random order. The subjects in this experiment are considered experts. Their relevancy ratings for each cluster are considered perfect. Table 10 shows an example of a part of a document that is sent user for relevancy rating.

Table 10. Example of relevancy rating document

Please Rate Clusters in scale of 1-7 . 1 is LESS RELEVENT with your interest – 7 is MOST RELEVENT with your interest and label the clusters if you can. Write your answers in the box below of the cluster.			
firebug firefox oper core norganna dublin wowinterfac viewer addon map download	apach leopard oreillimedia manag tomcat instal mac	tab grow back bobbi chord sunni door reviv creedenc clearwat beatl hebb lighthous guitar accur chordi lookin lyric edison	object learn identifi rdf practic xacml implement rfc gener metadata standard syntax content uniform guid resourc tabl framework best ieee uri sun
5 firefox addons	6 tomcat	6 music, guitar tabs	1 doesn’t ring a bell

6 watching video online	5 webserver on mac	6 ajax and web programing	6 watching video online

5.2 DISCUSSIONS OF RESEARCH RESULTS

5.2.1 Comparison of User Rated and System Predicted Cluster Set Qualities

In this section, Cluster Set Relevancy Score that is obtained after user relevancy ratings and Cluster Set Quality Scores that is implemented to predict the human judgment of the quality of interest clusters are examined whether the system prediction is successful determining the best clustering level.

Table 11 shows the scores after users' relevancy rating (CSRS) and Table 12 shows the system predicted scores for each clustering level (CSQS).

Table 11. Cluster Set Relevancy Scores after user relevancy rating (CSRS)

User ID	Number of bookmarks	5-cluster set	7-cluster set	9-cluster set	11-cluster set
1	88	4.00	4.00	3.10	2.90
2	104	2.20	2.71	2.33	3.82
3	113	3.60	2.00	1.88	2.50
4	185	4.20	4.00	4.78	3.36
5	243	2.80	2.57	2.30	2.18
6	281	4.80	4.86	4.00	4.27
7	303	5.80	4.71	5.56	5.73
8	325	4.40	4.50	4.60	4.70
9	358	3.60	4.29	2.56	1.91

Table 12. Cluster Set Quality Score that System Predicts (CSQS)

User ID	Number of bookmarks	5-cluster set	7-cluster set	9-cluster set	11-cluster set
1	88	0.22	0.12	0.08	0.05
2	104	0.46	0.04	0.24	0.12
3	113	-0.64	-0.01	0.05	0.03
4	185	0.04	0.25	0.16	0.17
5	243	0.15	0.22	0.07	0.12
6	281	0.66	0.62	0.55	0.62
7	303	-0.23	-0.07	-0.02	-0.05
8	325	0.22	0.11	0.19	0.12
9	358	0.35	0.23	0.03	0.06

Figure 11 shows the CSQS distribution on Number of Bookmarks.

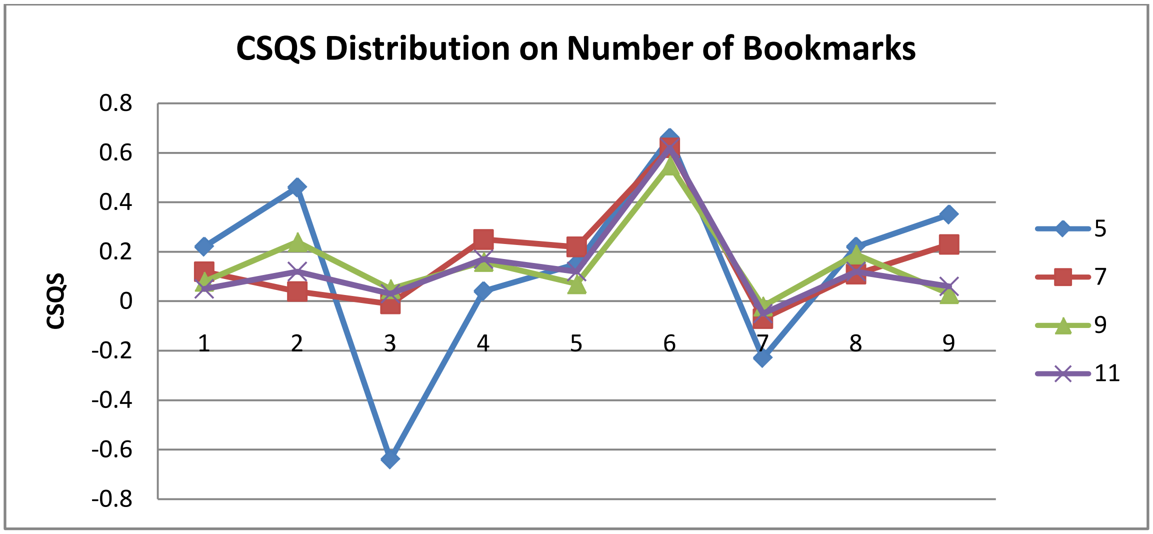


Figure 11. CSQS Distribution on Number of Bookmarks

Before comparing CSRS and CSQS, CSRS is normalized in order to have scores smaller than 1. Table 13 shows the normalized CSRSs.

Table 13. Normalized CSRSs

User ID	Number of bookmarks	5-cluster set	7-cluster set	9-cluster set	11-cluster set
1	88	0.57	0.57	0.44	0.41
2	104	0.31	0.39	0.33	0.55
3	113	0.51	0.29	0.27	0.36
4	185	0.60	0.57	0.68	0.48
5	243	0.40	0.37	0.33	0.31
6	281	0.69	0.69	0.57	0.61
7	303	0.83	0.67	0.79	0.82
8	325	0.63	0.64	0.66	0.67
9	358	0.51	0.61	0.37	0.27

Figure 12 shows the Normalized CSRS distribution on Number of Bookmarks.

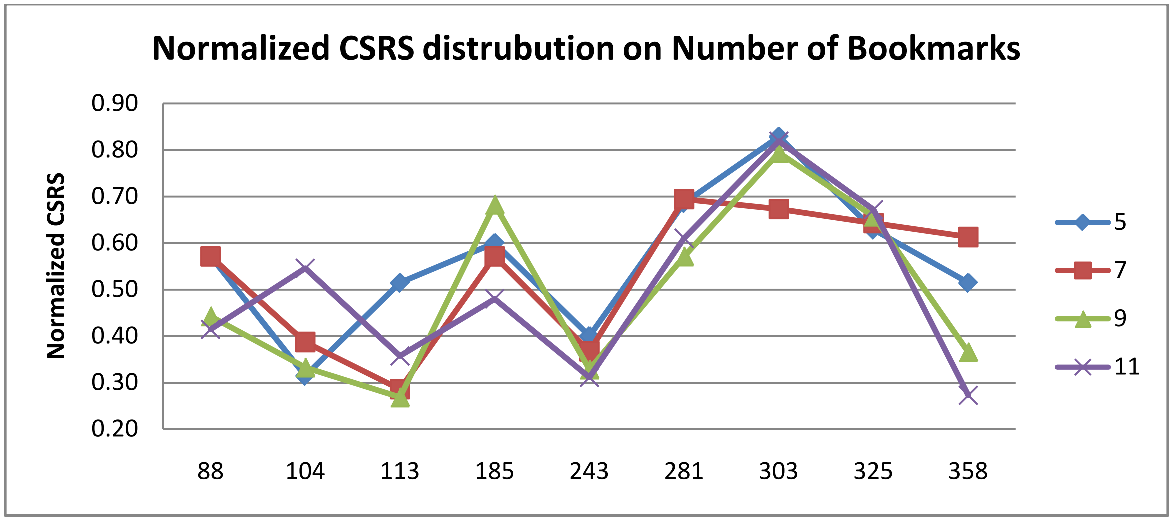


Figure 12. Normalized CSRS distribution on Number of Bookmarks

5.2.1.1 Analysis for the 2 Best Clustering Level

Normalized CSRS and CSQS are compared in order to see how system predicted score is close to the human judged scores for the 2 best clustering level. Table 14 shows system's two bests and users' two best cluster set levels with user relevancy scores and system predicted scores.

Table 14. Systems Two Bests and Users Two Best Cluster Set Levels and Scores

User ID	Number of bookmarks	System’s Best		System’s 2. Best		Users’ Best		Users’ 2.Best	
		Level	Score	Level	Score	Level	Score	Level	Score
1	88	5	0.22	7	0.12	5	0.57	7	0.57
2	104	9	0.46	5	0.24	11	0.55	7	0.39
3	113	11	0.05	9	0.03	5	0.51	11	0.36
4	185	11	0.25	7	0.17	9	0.68	5	0.60
5	243	7	0.22	5	0.15	5	0.4	7	0.37
6	281	7	0.66	5	0.62	5 ,7	0.69	5,7	0.69
7	303	11	-0.02	9	-0.05	5	0.83	11	0.82
8	325	9	0.22	5	0.19	11	0.67	9	0.66
9	358	5	0.35	7	0.23	7	0.61	5	0.51

Figure 13 shows the CSQS and CSRS scores for 2 best cluster set levels. CSRS1 and CSQS1 indicate the scores of the best levels. CSRS2 and CSQS2 indicate the scores of the second best levels.

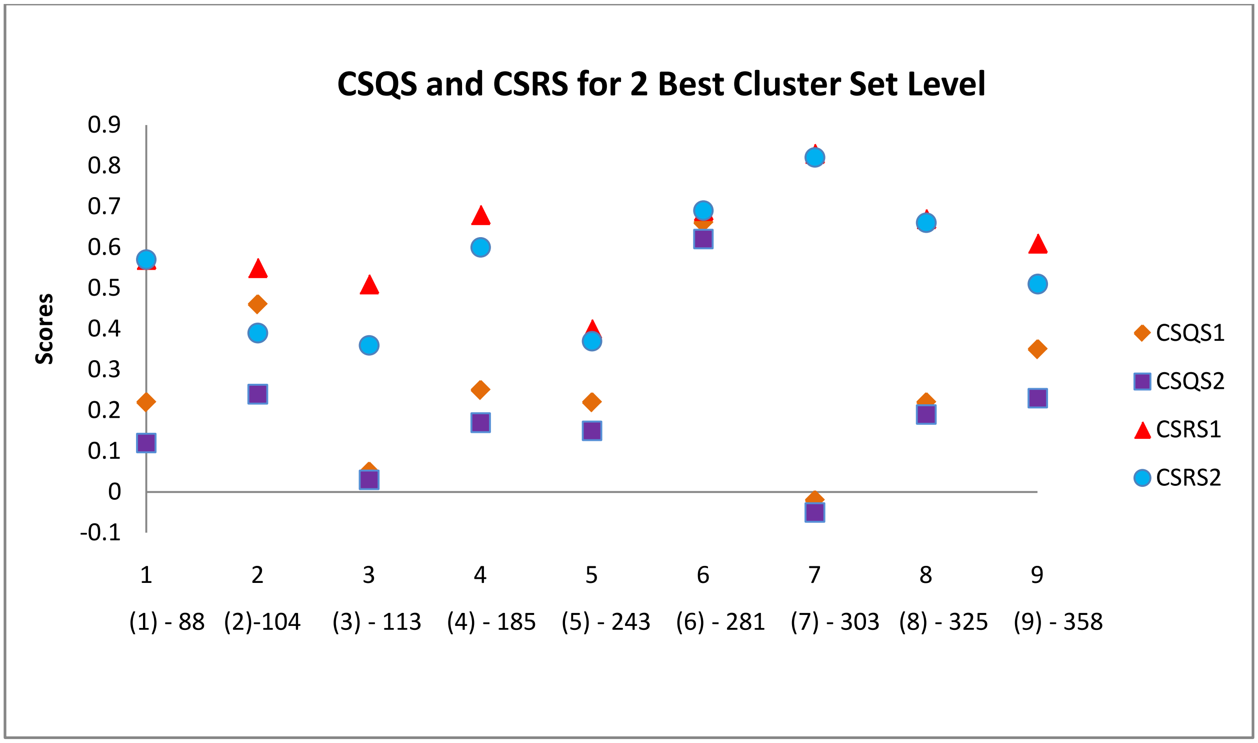


Figure 13. CSQS and CSRS scores for 2 Best Cluster Set Levels

Figure 14 shows the System's Best and Users' Best Cluster Set levels and how they fit.

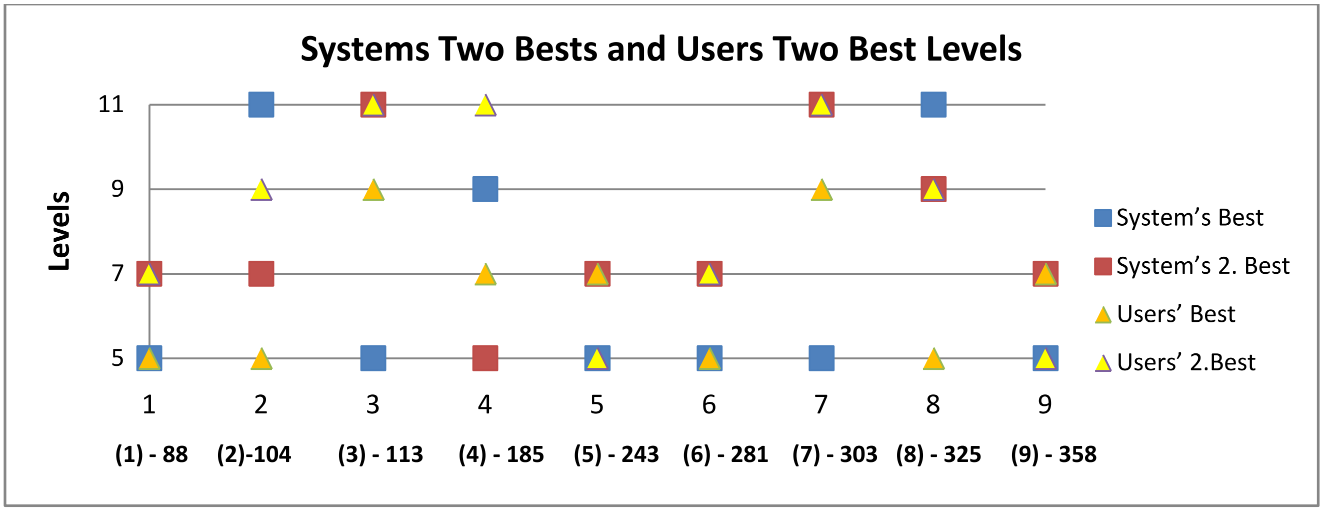


Figure 14. System's Best and Users' Best Cluster Set level

It would be perfect if all user's best matches with system's best and user's second best matches with system's second best. But there are some cases such as user's best matches with system's second best or user's second best matches with system's best or none of them matches at all. In order to see the differences between the probabilities of being accurate or not, all probabilities are scored from 0 to 4:

- 4 is when system's best and users' bests are all different.
- 3 is when system predicts only the best or the second best with wrong order.
- 2 is when system predicts only the best or the second best with correct order.
- 1 is when system predicts both users' 2 bests with wrong order.
- 0 is when system predicts both users' 2 best with correct order.

Figure 15 shows the prediction difference score.

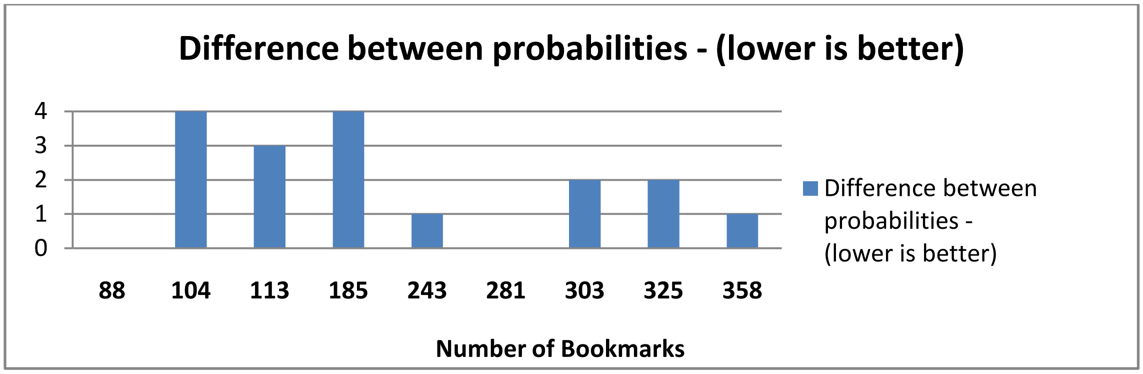


Figure 15. Prediction Difference Scores

It looks like system prediction is better with less than 100 bookmarks and if number of bookmarks between 200 and 300. With more than 300 bookmarks system prediction does not look good. In order to draw a conclusion, more data is needed.

5.2.2 Effect of the Number of Bookmarks and Number of Folders of Bookmark set on CSRS

After getting user relevancy ratings, we looked at two things:

1. Relationship between total number of bookmarks in a bookmark set and user satisfaction of created interest clusters for each clustering level.
2. Relationship between number of folders in a bookmark set and user satisfaction of created interest clusters for each clustering level.

Figure 16 shows the distribution of CSRS on number of bookmarks. As it is seen, none of the clustering levels has a linear relationship by the increasing order of number of bookmarks.

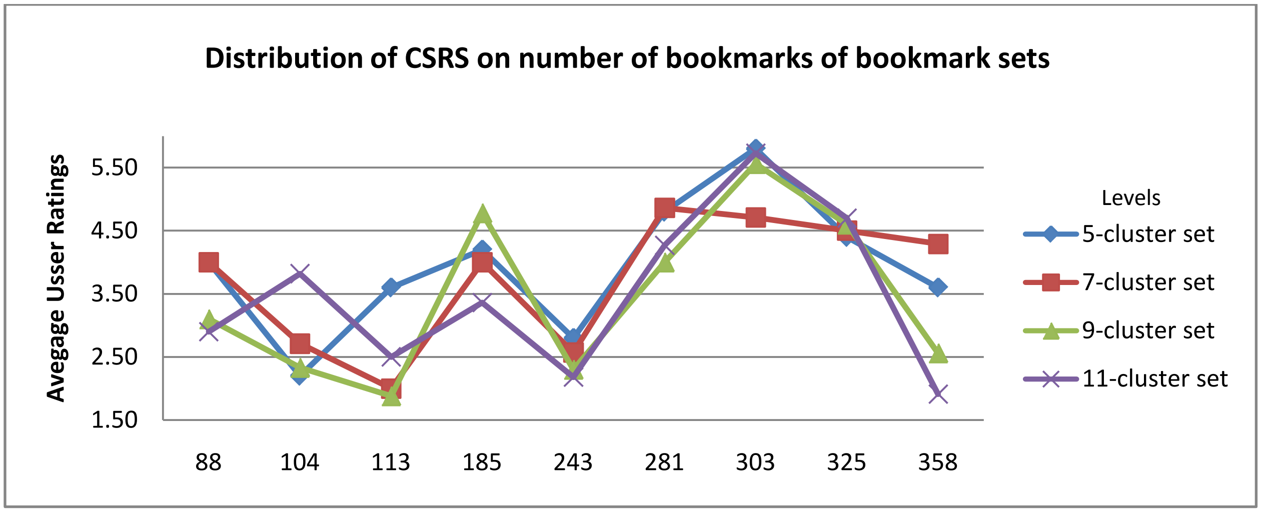


Figure 16. Distribution of CSRS on number of bookmarks

As the same above, there is no relationship between CSRS and the increasing number of folders of bookmark sets as seen on Figure 17.

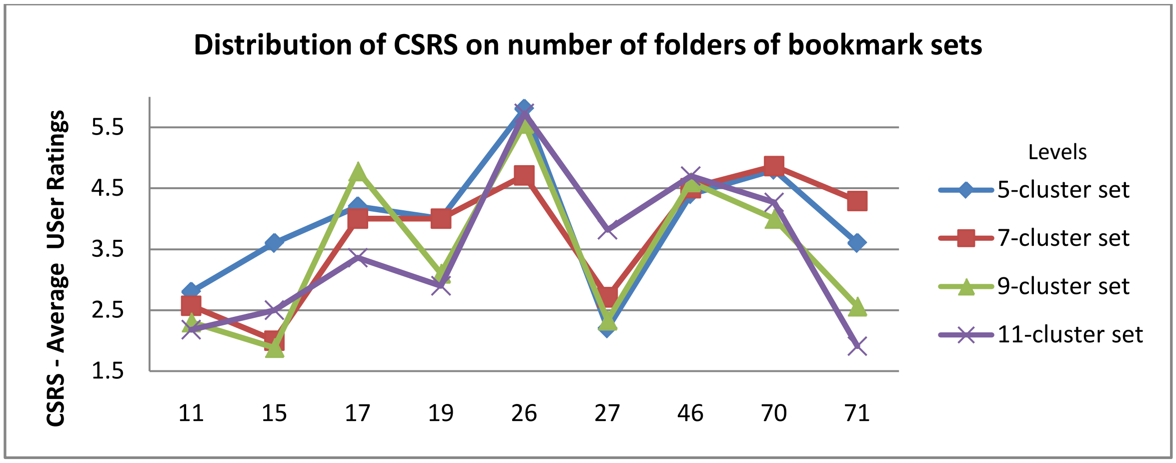


Figure 17. Distribution of CSRS on number of folders of bookmark sets

6.0 CONCLUSIONS

This chapter describes the contributions and implications of this research. Then, future work is discussed.

6.1 CONTRIBUTIONS AND IMPLICATIONS

We examined whether a user's bookmarks can be used as a source of information that will reflect a user's interests. Ways to extract the best descriptive information from bookmarks were explored. A bookmark is represented with all body text including bookmark title and also represented just with bookmark title and H1 text in the body. It was found that a combination of H1 text and bookmark titles provided the best descriptive information for a bookmark.

After each bookmark is replaced with its representation, they are all converted to vectors so that a bookmark set becomes a bag of bookmark vectors. K-Means clustering was performed on a bookmark set of create four sets of clusters - 5, 7, 9, 11 cluster sets. The better quality clusters a cluster set has, the clustering level returns a better estimation of user interest than other levels. A cluster is a list of terms extracted from bookmarks. To measure the quality of clusters, all clusters are converted into tag clouds to present visualized representation of clusters. These visualizations were displayed randomly to users for relevancy rating. After user ratings, for each clustering level average of rating scores are calculated as cluster set relevance score (CSRS) and it is used to determine the best clustering level that is an indication of better representation of user interest.

CSQS equation is implemented to predict the chances of levels being the best without user interaction looking at the number of bookmarks in the clusters that falls into the same folder. It appears that CSQS works for predicting best two clustering levels with less than 100 bookmarks and if number of bookmarks between 200 and 300. With more than 300 bookmarks system seems like not working well. To draw a solid conclusion it is necessary to conduct the study with larger data set.

6.2 FUTURE WORK

This study explored one way to determine the best clustering level for bookmark information in order to get most relevant clusters for user interest. It is necessary to conduct a large and more controlled study of bookmarks.

Browser bookmarks are used in this study. Current trend is online bookmarking sites and similar study might be conducted using delicious online bookmarking site as basis. Unlike from browser bookmarks, tags are used for organization of URLs. Usage of tags might cure some problems that occur in browser bookmarks. For example, the monolithic structuring of URLs that means a URL can be listed in one folder without duplicates in browser bookmarks. Since a resource can be tagged with several tags, online bookmarking sites overcome this problem. Tagging system also gives user a new way to conceptualize resources better than browser bookmarks' hierarchical folder structure.

In our study only bookmarks are used as user interest source. Since user interest is more dynamic than bookmarks, other implicitly aggregated information sources might be integrated with bookmarks to construct a profile – eg. monitoring a user's browsing history.

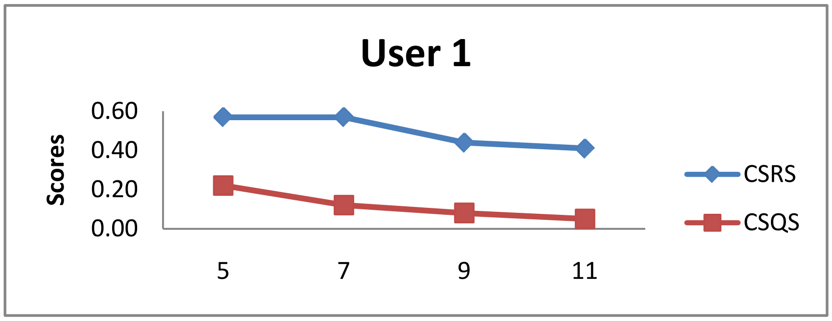
Our study is part of a bigger research question: Ranking RSS feeds using bookmark based user profiles. For the future work, an RSS reader could be implemented that takes user bookmarks to create user profile and displays a ranked list of RSS feed entries based on bookmark user profile. As literature shows, most of the current RSS readers do not provide personalization. If they provide personalization, they use either explicitly provided keyword by user as user profile or large history, browsing data that are computationally expensive to mine. There has not explored a new way to make use of RSS readers to create user profile without user interaction but economically constructed.

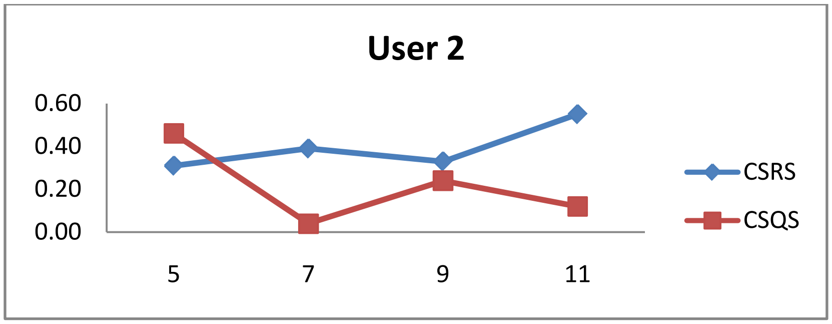
The goal of this future research is to determine if the user bookmark information could be used to improve user experience by providing a better presentation order of feed entries and ranking of entries compared to other news readers. Dr. Michael Spring, advisory of this thesis, has an implemented RSS aggregator that used k-means clustering techniques to build clusters of news. These clusters might be mined to determine user points of interest looking at the users' news access patterns. Since RSS feeds are very dynamic, bookmark information might be combined with user access pattern to predict user interested news. Envisioned system over view is that user provides his/her bookmarks for user profile construction and RSS feeds that user reads regularly. These feeds are saved into the RSS reader and as user uses the reader, it collects user news access pattern and combine this information with created bookmark based user profile and returns a ranked list of feed entries for user relevancy rating.

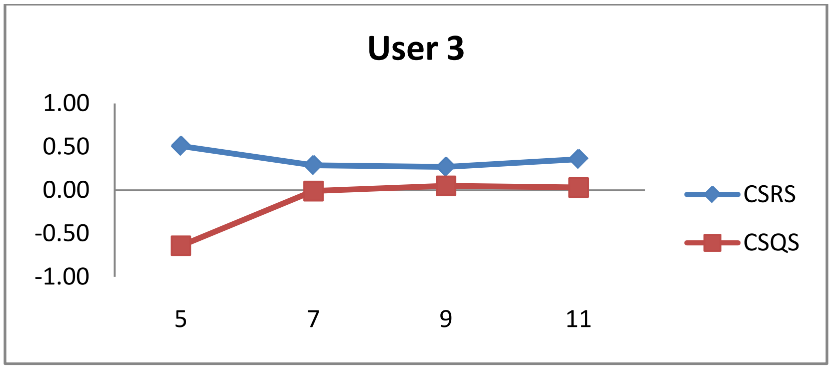
$$NDCG_q = M_q \sum_{j=i}^{K} \frac{(2^{r(j)} - 1)}{\log(1 + j)}$$

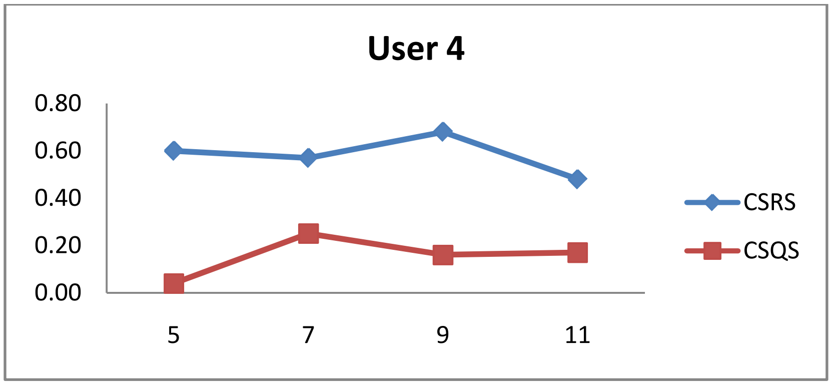
where M_q is the calculated normalization constant. The perfect ordering would obtain NDCG of 1; each r(j) is an integer representing the relevancy rated by human. NDCG rewards relevant documents in the top ranked results more heavily than those ranked lower and punishes irrelevant documents by reducing their contributions to NDCG.

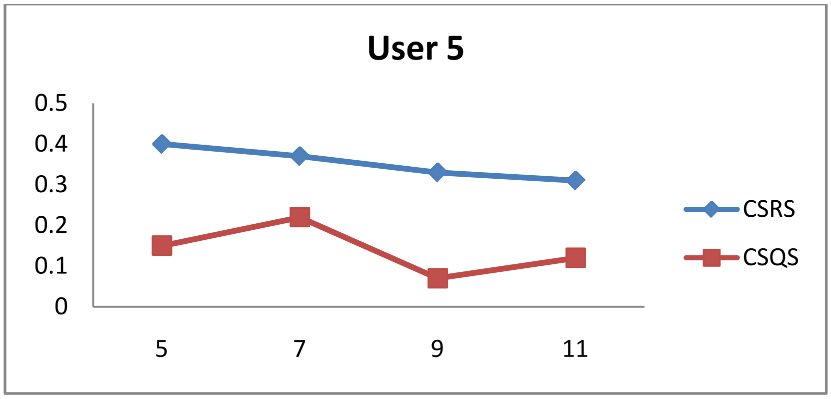
7.0 APPENDIX A. CSRS AND CSQS COMPARISON CHARTS OF SUBJECTS

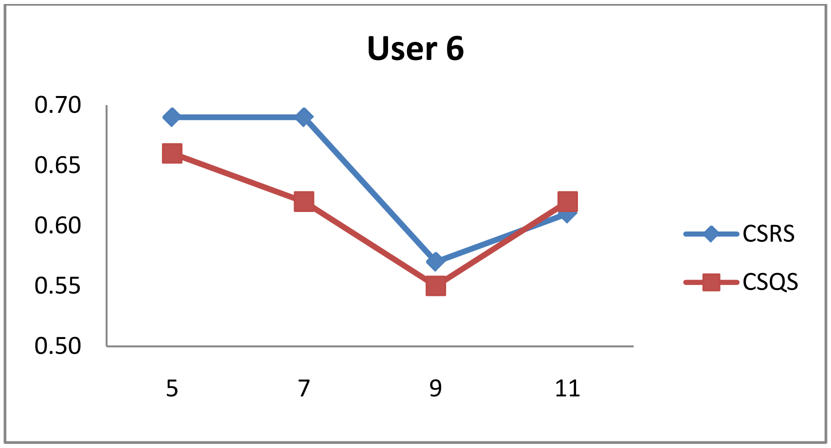


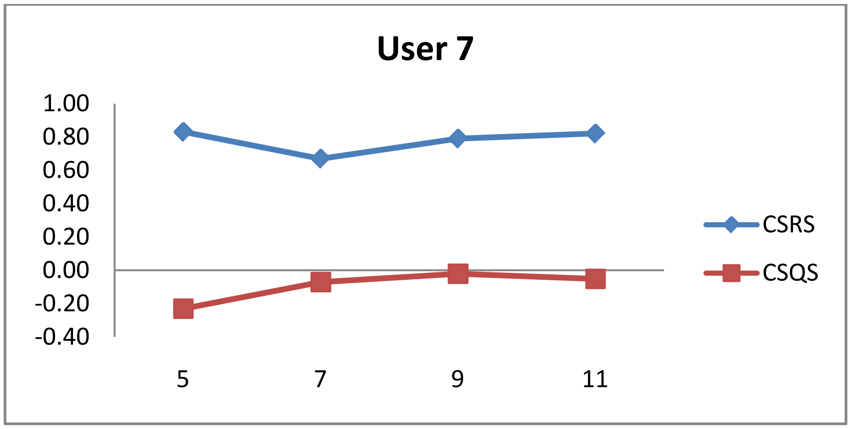


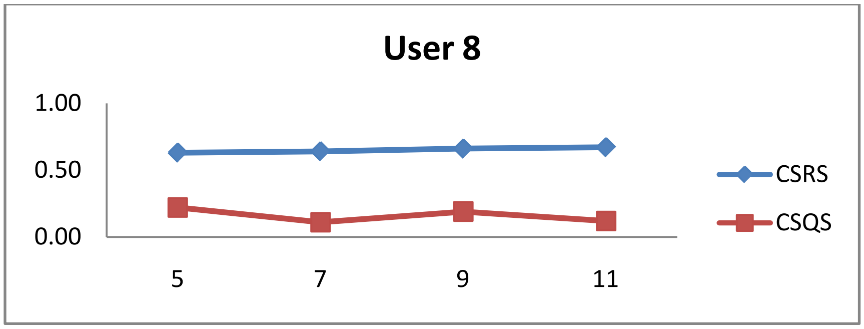


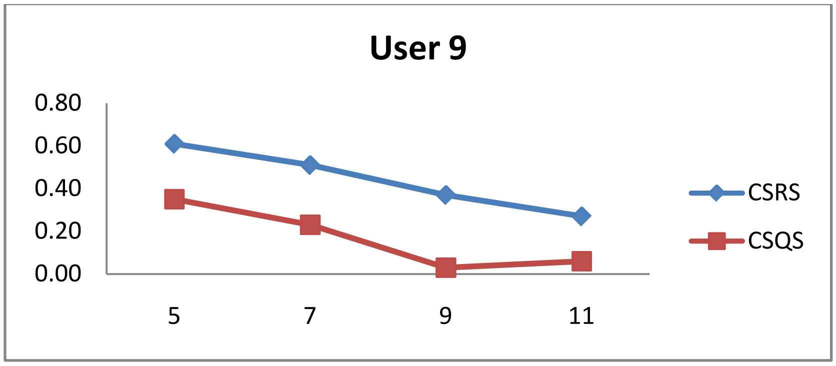



User 5
0.5
0.4
0.3
0.2
0.1
0
5
7
9
11
CSRS
CSQS


User 6
0.70
0.65
0.60
0.55
0.50
5
7
9
11
CSRS
CSQS


User 7
1.00
0.80
0.60
0.40
0.20
0.00
-0.20
-0.40
5
7
9
11
CSRS
CSQS


User 8
1.00
0.50
0.00
5
7
9
11
CSRS
CSQS


User 9
0.80
0.60
0.40
0.20
0.00
5
7
9
11
CSRS
CSQS

8.0 BIBLIOGRAPHY

Bharat, K., Kamba, T., & Albers, M. (1998). Personalized, interactive news on the Web. *Multimedia Systems* , 349-358.

Billsus, D., & Pazzani, M. (2000). User Modeling for Adaptive News Access. *User Modeling and User-Adapted Interaction* , 147-180.

Bloglines . (n.d.). Retrieved January 2010, from Bloglines : www.bloglines.com

Budzik, J., & Hammond, K. J. (2000). User interactions with everyday applications as context for just-in-time information access. *Proceedings of the 5th international conference on Intelligent user interfaces*, (pp. 44-51).

Carreira, R., Crato, J., Goncalves, D., & Jorge, J. A. (2004). Evaluating adaptive user profiles for news classification. *Proceedings of the 9th international conference on Intelligent user interfaces*, (pp. 206 - 212).

Catone, J. (2009, December). *10 Web Tech Innovations That Have Improved Our Lives*. Retrieved December 2009, from 10 Web Tech Innovations That Have Improved Our Lives: http://mashable.com/2009/12/14/web-tech-innovations/

Chan, C.-H., Sun, A., & Lim, E.-P. (2001). Automated Online News Classification with Personalization. *Proceedings of the 4th International COnference of asian Digital Library*, (pp. 320-329).

Chan, P. K. (2000). Constructing Web User Profiles: A Non-invasive Learning Approach. In *Web Usage Analysis and User Profiling* (pp. 39-55).

Chen, L., & Sycara, K. (1998). WebMate: a personal agent for browsing and searching. *Proceedings of the second international conference on Autonomous agents*, (pp. 132-138).

Chien, C. C., Chen, M. C., & Yeali, S. (2004). PVA: A Self-Adaptive Personal View Agent. *Journal of Intelligent Information Systems* , 173-194.

Chirita, P.-A., Olmedill, D., & Nejdl, W. (2004). PROS: A Personalized Ranking Platform for Web Search. In *Adaptive Hypermedia and Adaptive Web-Based Systems* (pp. 34-43).

Conner, H. J. (1967). Selective Dissemination of Information: A Review of the Literature and the Issues.

Crabtree, B., & Soltysiak, S. J. (1998). Identifying and tracking changing interests. *International Journal on Digital Libraries* , 33-38.

Del Corso, M. G., Gulli, A., & Romani, F. (2005). Ranking a stream of news. *Proceedings of the 14th international conference on World Wide Web* (pp. 97-106). ACM.

Google Reader . (n.d.). Retrieved January 2010, from Google Reader : reader.google.com

Howley, M. (2009, 11 2). *Communities of Practice: Optimizing Internal Knowledge Sharing*. Retrieved from UXmatters: http://www.uxmatters.com/mt/archives/2009/11/communities-of-practice-optimizing-internal-knowledge-sharing.php

Hyoung-Rae, K., & Chan, P. K. (2008). Learning implicit user interest hierarchy for context in personalization. *Applied Intelligence* , 153-166.

Jung, J. J., & Jo, G.-S. (2003). Extracting User Interests from Bookmarks on the Web. In *Advances in Knowledge Discovery and Data Mining* (p. 568).

Kiran Jude Fernandes, V. R. (2005). Portals as knowledge repositories and tranfer tool - VIZNCon case study. *Technovation 25* .

Li, W., Vu, Q., Agrawal, D., Hara, Y., & Takano, H. (1999). PowerBookmarks: a system for personalizable Web information organization, sharing, and management. *Proceedings of the 1999 ACM SIGMOD international conference on Management of data,* (pp. 565 - 567).

Nakajima, S., Oyama, S., Sumiya, K., & Tanaka, K. (2002). Context-dependent Web bookmarks and their usage as queries. *Processings of the 3rd International Conference on Web Information Systems Engineering.*

Opfine - Opinionated Financial News. (n.d.). Retrieved January 2010, from Opfine - Opinionated Financial News: http://www.opfine.com/

Pazzani, M., & Billsus, D. (1997). Learning and Revising User Profiles: The Identification of Interesting Web Sites. *Machine Learning* , 313-331.

Pon, R., Cárdenas, A., Buttler, D., & Critchlow, T. (2007). iScore: Measuring the interestingness of articles in a limited user enviroment. *IEEE Symposium on Computational Intelligence and Data Mining 2007*.

Pon, R., Cárdenas, A., Buttler, D., & Critchlow, T. (2007). Tracking multiple topics for finding interesting articles. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 590-569). ACM.

Regator . (n.d.). Retrieved January 2010, from Regator : www.regator.com

Starr, B., Ackerman, S. M., & Pazzani, M. (1996). Do-I-Care: a collaborative Web agent. *Conference companion on Human factors in computing systems: common ground*, (pp. 273 - 274).

Techmeme. (n.d.). Retrieved January 2010, from Techmeme: http://www.techmeme.com/

Thomas H. Davenport, D. W. (1998). Succesful Knowledge Manangement Projects. *Managing the Knowledge of the Organization* .

Trajkova, J., & Gauch, S. (2004). Improving ontology-based user profiles. *Proceedings of RIAO*.

Wikipedia. (n.d.). Retrieved January 2010, from Wikipedia: http://en.wikipedia.org/wiki/RSS