A different approach for an URL recommendation system

Vu Nam Nguyen
Student I.D. 1842609

VU Bachelor Project, Informatie, Media & Management
(*under* the *supervision of Victor de Boer)*

vnn230@vu.nl

Abstract

In this paper, we analyze several types of search engines and similar website tools. The current similar website tools return URLs using other methods than user generated content. The internet has seen an increase in user generated content. The classifying of content based on social tagging might help similar website search return better results. We assume a tool that presents similar URLs based on social tagging would give back better results than the existing similar website tools. We build a search tool that can compare URLs and return similar websites based on user generated content. We analyze the best method to search our database when using the TF/IDF search algorithm and analyze our usability options for creating the tool. The tool is then evaluated for usability and compared with other similar website tools. A link to the tool can be found here http://websfav.com/

1 Introduction

Search engines like Google, Yahoo & Bing are major forces in today's information driven society. These search engines handle millions of search queries each day and have comprehensive indexes of the World Wide Web. As the web grew and still continues to grow, these search engines have scaled proportionally to the amount of available documents. Also, the web further develops into more social areas due to large internet communities such as Twitter, Facebook, Hyves and so on. Thus creating more user generated content on the internet. With this type of content, users are able to decide the importance and classification of a link, image or article. Many search engines exist with each a different purpose. From search engines that goes through products of an ecommerce website or search engines that go through the whole connected web. However, there is still a lack of an URL recommendation search tool with the sole purpose of harnessing social data and returning a list of interesting and relevant URLs back to the user. How can we integrate social media for better search results according to user generated data? In what ways can user generated content enhance traditional methods of information retrieval in website similarity? In this paper, we will try to find an answer to these questions by building an URL recommendation system based on user generated content, and through analyses.

2 The need for social content in search engines

In the following section, an overview of what search engines do will be given. Furthermore, the growing importance of user content will be explained and why it should play a bigger role in web searches.

2.1 Traditional search engines

In the past, information on the internet used to be static, consisting of countless pages with information and images available for internet surfers to browse on. They go through millions of pages and present you with these static pages that match your search criteria. The URLs are gathered by webcrawlers, which are automated bots that visit link to link and archive the data of a website. Then ranking algorithms decide which of these URLs are most relevant for the given search queries. Google, for instance, calculates the importance of a website by having each page cast a vote for another page through hyperlinks. Thus, a website's relevancy and ranking is decided by other web documents. Due to this, newer web pages with the same contextual matching do not threaten the ranking of older more established websites (Avrachenkov et al., 2004).

2.2 A social web

Users slowly gave a higher importance on usability and sharing content. The key for this is user participation (Tim O'reilly, 2005). As a result of information traveling between users worldwide; tagging, social content, social networking and online communities were established. With the entrance of user participation, people can manipulate and share data instead of only retrieving it. A study in 2007 has shown that if user generated content from social bookmarking websites continues to grow, it can outgrow the scale of the internet from 2007 within a few years time (Heymann et al., 2007). Search engines should therefore focus more on indexing social content instead of static content. More

specifically, more importance should be given to the indexing of user images, social tagging, user comments, and so on. For example, user tagging can be seen as a description in several words created by one user. So if multiple users provide tags for an URL then the more accurate the 'description' will be. The overlapping tags determine the top tags for a website and the tags that are irrelevant will fade out and become less important (Heymann et al., 2007).

We assume that this method of social tagging can increase relevancy for similar websites. If each website contains tags added by a user then other websites that are also tagged with similar tags can be recommended back to other users.

3 The website recommendation tool

The main part of our research is to build a system that can return similar websites when given an URL based on social tagging. To find out what our tool requires we have made an analysis between the different social bookmarking websites, current website recommendation tools and search engines.

3.1 Analyzing different website types

We differentiated between the following website types because their search has similar components such as user tags, user content, searching URLs etc. We came up with the following types: blogs, social community, social bookmarking, web directory, search engines and similar website tools. The topics that we have analyzed are based on the search function of these websites. We have asked ourselves if the search uses either user content, indexed URLs, tags, the text of a document or offer to compare URLs (Appendix A).

Table 1: Results of the analysis of different search systems

Type	Uses user content	Indexes URLs	Searches tags	Searches text	Compares urls
Blogs	Yes	No	Yes	Yes	No
Social communities	Yes	No	Yes	No	No
Social bookmarking	Yes	No	Yes	No	No
Web directories	No	Yes	No	No	No
Search engines	No	Yes	Yes	Yes	Yes
Similar website search	No	Yes	No	No	Yes
Our recommend tool	Yes	Yes	Yes	No	Yes

We found that a similar website search with informational and relevant results is still lacking. The main problem with the current similar website services is that these websites are not accurate enough. It's not clear for the user what the results are based upon. They simply see a list of URLs and have to try to review each URL themselves.

3.2 Requirements website recommendation tool

When setting up the requirements for our tool, we used the analysis of different search systems and a few user scenarios that the tool needs to satisfy (Appendix B).

Table 2 below shows an overview of the requirements for similar website search tools that were built along with their motivation.

Table 2: Tool requirements

Requirement No.	Type	Motivation
1	Search Box	The search is the main component of the tool. Users must be able to type in their search query.
2	URL search	Users can search by entering an URL
3	Keyword search	Users can search by entering tags / keywords
4	Smart search	The tool can recognize if the user search for an URL or tag(s)
5	Ranking	The tool must use a ranking system to differentiate between results
6	Multiple sort options	The search returns a ranked list based on popularity or relevancy
7	iFrame	The tool offers the user to view the URL in a frame so that the user remains on the website
8	User reviews	The tool displays reviews of the URL so the user can make better choices
9	Performance	The tool must be accurate and short search times to satisfy performance needs

4 Usability and design choices

Usability of a website refers to how well the user interaction is within a website. Websites with good usability allow users to quickly and easily browse a website. Finding information is of high importance. If users have to search hard for the information they're a looking for, they might leave the website. A good website makes the visitor stay and enjoy the contents of the website. The needs of a user should be considered when planning a website.

4.1 How do users think?

Users like quality content over web page design. Poorly designed websites have proven to gain visitor traffic with quality content. Users visit a page for the content it provides and not for design. Users don't read a whole web page. They analyze it by quickly scanning a web page. They search for recognizable patterns in design and use these as fixed points or anchors to guide them through the content. Searching for website functions are considered bad. If a user has to put too much effort into finding links or buttons than he/she will leave the website (Friedman 2008).Therefore recognizable and structured web pages are considered best.

4.2 Usability layout

We think that a clear and well defined layout is very important for the usability. We used as many guidelines (U.S. dept. of health and human services, 2006) for the layout as possible and found that our tool should include the following:

- Users are able to easily compare and analyze search items to discern similarities, differences, trends and relationships.

- Searches are first ranked from high importance to low.
- Page elements are placed in a grid for easy browsing.
- The layout is fluid and able to adapt to various screen sizes.
- Larger screens see more results aligned and so on.
- Blank space is also limited so that more of the canvas can be used for information.
- The search input is placed in the middle and on top. This is a good place to put the search since this is the main interaction point for the user.

When building the tool, several design choices and guidelines became more important than others. We assume that these choices will be of great value for the purpose of our search tool.

Table 3: Unique design choices

Design choice No.	Type	Description
1	Stay on site	When visiting another website the user sees the website in an iFrame. The user has the option to leave the website but is encouraged to remain on the search website to quickly try out another link.
2	Performance	To avoid slow loading times and quick view for the user, we have chosen to use AJAX so that a page refresh and load is not needed. The user types in keywords and can see the results immediately on screen.
3	Compact information	To give as much information as possible in a small area, results come back on cards. Since not all users are interested in every URL certain information should be only be accessed upon request. The cards can be flipped to see the reviews of a certain URL.
4	Visual effects	We use some jQuery effects. It is more pleasing to the eye when the loading of new content fades in slow instead of jumping suddenly on screen.
5	Preview	Images are helpful visual tools to attract the attention of the user. Visual content can also help a user determine if he/she likes the URL or not.

5 Database and algorithm design

The tool will need a database to store all the URLs and tags. In this section we will go through the structure of the database and the dataset we used for the project.

5.1 The delicious folksonomy

The folksonomy of the social bookmarking website Delicious.com is freely available and offers a rich dataset of user-generated classification of URLs. For this reason we decided to use the delicious folksonomy from 2009, since this contains all the data to index our URLs with already defined user tags.

5.2 Search and ranking algorithm

The search can return popular and relevant results. For the relevant results a tf x idf scoring function is used. We chose this algorithm because we were working with MySQL and this algorithm is build in so we can use the algorithm in real time with minimal performance loss.

$$s_{\mathrm{TF/IDF}}(d_j, \mathbf{Q}) = \sum_{i=1}^{m} w_{ij} \ln\left(\frac{N}{c_i}\right)$$

A term can be compared with a tag for the document we use all the tags of an URL. When matched, it searches the whole text for the search terms. This scoring function scores a document by the sum of the weights of the query terms it contains. We use the tags as query terms and match that against the tags of another URL.

5.3 Extract data

To extract the data needed to fill the database we wrote a program in Java to open and extract the content from the dataset. It extracts each URL along with their tags and also keeps count of each tag. The tags are then ranked based upon how many users have tagged the URL with the same keywords. This helps to determine which tags are more important to a website. We came up with two methods regarding the search and the design of the database. We tested these methods on performance and relevancy.

Method 1: Use all the tags from every user for every URL and store them in a table.
Method 2: Match the search tags only for the most popular tags.

Hypothesis: This first method will give the most accurate results while the other will show better performance.

5.4 Testing method 2 for best N top tags

Imagine an URL being tagged 4000 times. Most of these 4000 tags are the same words and thus overlap each other. The words that are tagged the most are called the top tags. To test the hypothesis, we need to define the best number of top tags for method 2.

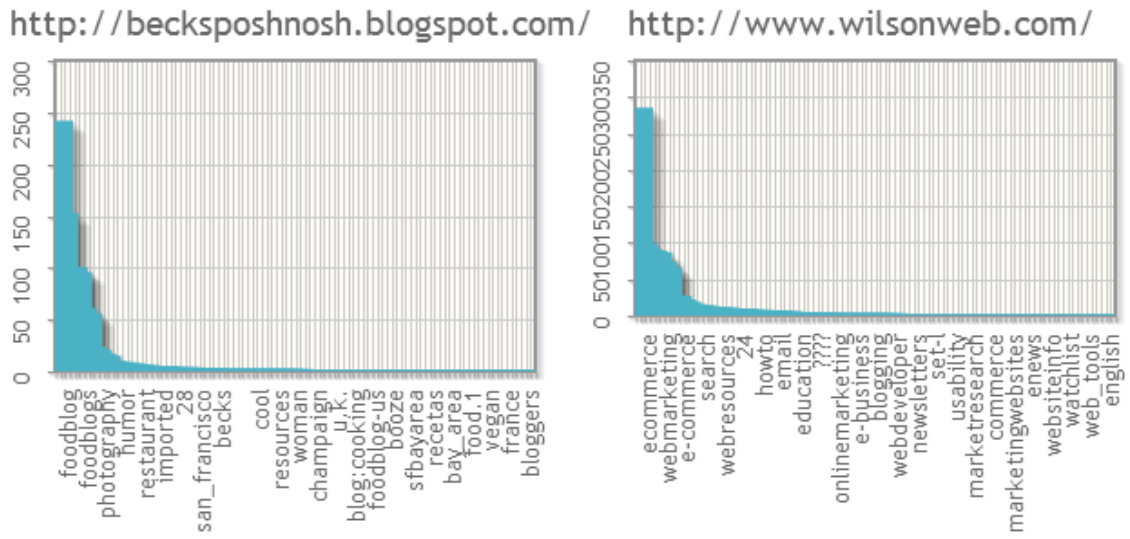

Figure 1: An example of 2 bar charts showing the distribution of top tags. We can see the charts show a very unbalanced histogram of the top tags resulting in a very long tail.

Table 4: N Top tags overlapping percentage of all tags

	N = 5	N = 10	N = 15	N = 20	N =25	N = 97
Area Top Tags	57,6%	68,4%	73,4%	78,9%	80,3%	95,1%

We used the Central Limit Theorem (Efstathiou) and took a sample of 30 out of 12.615 URLs and placed them in several bar charts. On the x-axis we show the number of tags and on the y-axis the tagged keyword.

We started at 5 top tags and increased each iteration of our test with another 5. At 25 we stopped, when it became apparent that the first few top tags contained most of all the tags.

5.5 Overall performance test

For the performance test we used a single keyword search and an URL search with the heaviest search load using www.redmine.org.
This URL has 9188 tags with a total of 94883 chars (An A4 paper would fit about 1500-2000 charracters).

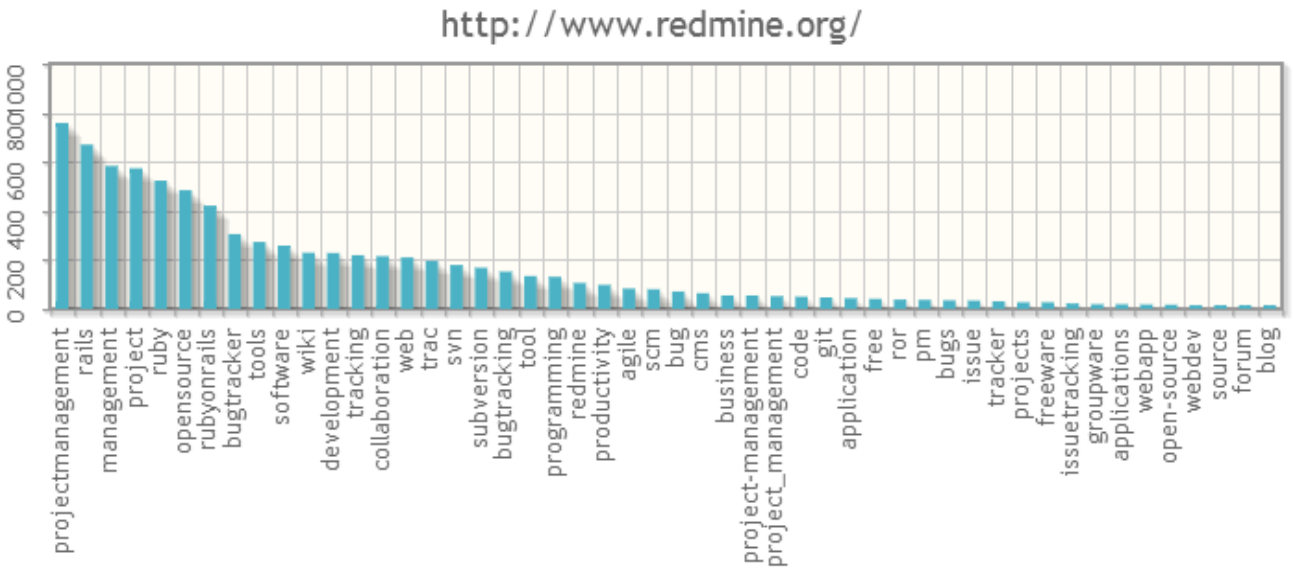

Figure 2: A bar chart of the top 50 most popular tags

The following tests are conducted in a simple 'query only' environment so loading of images and so on are not considered yet. Single or multiple tags are being matched with N tags located in the database. *N=All* is all the tags in the database, so matching all tags with *N=All* means every tag of the search URL is being matched with every tag of every other URL in the database. The *N=All* method supports count of individual tags while the N =5 to N = 25 are simplified and only hold the n top tags without count.

Table 5: Query times of different methods

Tags	N = 5	N = 10	N = 15	N = 20	N =25	N = All
Single tag (Rare occurrence)	0.0008 sec	0.0008 sec	0.0008 sec	0.0009 sec	0.0009 sec	0.0229 sec
Single tag	0.0323 sec	0.0326 sec	0.0334 sec	0.0353 sec	0.0368 sec	0.2756 sec

(High occurrence)						
Multiple tags (Rare occurrence)	0.0046 sec	0.0059 sec	0.0065 sec	0.0074 sec	0.0087 sec	0.5117 sec
Multiple tags (High occurrence)	0.0320 sec	0.0328 sec	0.0338 sec	0.0352 sec	0.0365 sec	0.3626 sec
All tags	0.0915 sec	0.1183 sec	0.1298 sec	0.1515 sec	0.1753 sec	1.3665 sec

**Single tag: ekonomika (low, 16 occurrences); blog(high, 4620 occurrences)*
**Multiple tags: party, space, chair(low, 66 occurrences); fun, cool, blog(high, 1787 occurrences)*

5.6 Results

We found that when testing the best N top tags for method 2 that when N=10, it contained 68.4% of all the tags. So the top 10 tags accumulate for 68.4% of the top tags which would be sufficient enough if a confidence interval of 68% is satisfying. Further testing has shown that on average 97 tags define 95.1% of all the tags. So for the best accuracy, any N more than N=10 would have many overlapping top tags.

The results on performance show that queries with a smaller N tag field perform better. We can see that the query times differ between N = All and N = 5 to N = 25. We assume that searching through 25 top tags is less heavy than searching through 9000 tags. Although the load times differ, we assume that the performance for *N=All* is still good enough for its purpose. Further research is needed but *N=All* can hold since the load times are within range of user satisfaction (Gomez, 2006). For heavier purposes the database should use N top tags where N is at least 10. During the performance tests it was also noticed that the more tags were used the more results were shown. Thus, the hypothesis is correct since the use of *N=All* is preferred over N top tags. In our case, method 1 is the best because the loading times are satisfactory for our purpose.

6 Tool report

We finished building tool and implemented all the features we wanted the tool to do. The tool is located at the following url: http://websfav.com.

6.1 Website functionality

To use the tool, a user fills in his search URL or keyword(s), *see fig. 3*. There is a filter that can rank the results according to relevancy or popularity. The results are then shown in a grid with 'cards', *see fig. 4*. Each card contains a screenshot, 5 top tags, amount of users that tagged the URL, a review button and a view website button. A user can read the reviews by clicking the review button. The card flips and the basic URL information is replaced by the reviews of the URL, *see fig. 5*. Clicking the info button returns the user to the information card. On the information card the user can click the top URL to begin a new search with the selected URL, the same goes for when the user clicks on the top tags. When a user clicks on the view website button a lightbox appears with the website of the URL in a frame. The user can browse the selected URL while remaining on the main website. If he/she chooses to view the website in a new tab, the/she can click on the link on the bottom left.

Figure 3 the index page where you can type in an URL or keyword(s).

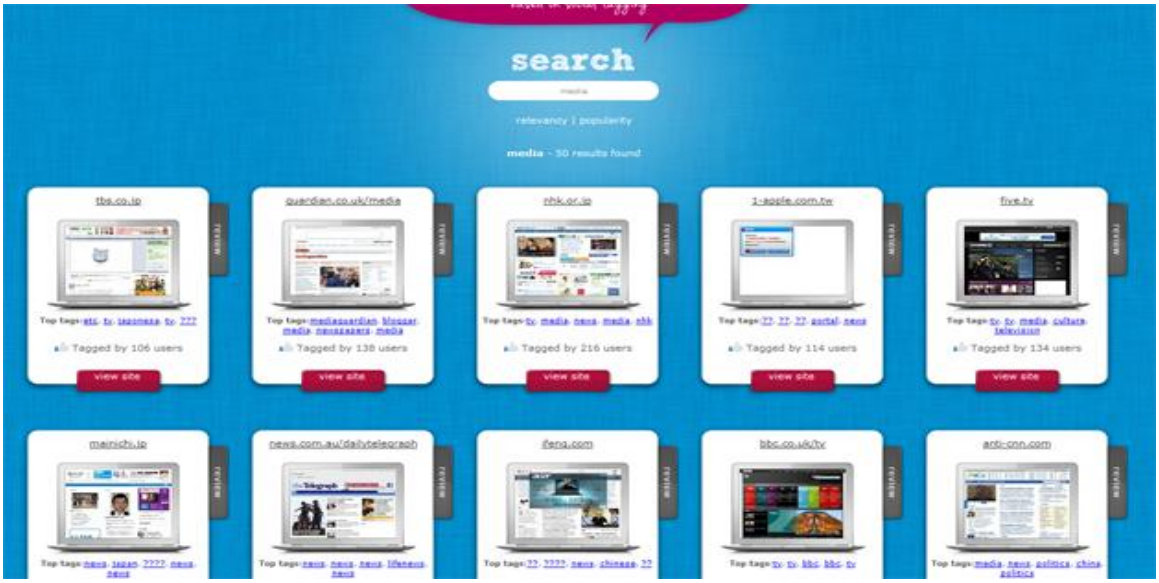

Figure 4: the result page where all the results are listed with a screenshot and tag information.

Figure 5: a close up of three results. Left is an example of website reviews and the other two contain URL and tag info.

7 Evaluation

According to Jacob Nielsen (Nielsen, J., 2012) a usability test can consist of only five subjects and three small tests. We conducted the first usability test with 5 users. The idea was to let people decide for themselves what the website does and how the website works. If the usability is good then it should be clear to the user. Each user spent 5 minutes browsing the website and was told to list every problem they came across. At the end they filled in a few likert scales to determine if the tool satisfied their expectations. We then asked them to review the two leading website similarity tools (http://www.similarsites.com, http://www.similarsitesearch.com) and fill in the same likert scales.

Table 6: Usability test

component	Question	Our Score	Similar Sites	Similar Site Search
Learnability	How easy was it to learn the basic tasks?	4,4	4,2	4,4
Efficiency	How quick did you perform the tasks?	3,8	3,6	3,4
Utility	Does the website do what you want it to do?	4,4	4,2	3,6
Errors	How many errors did you make? (1 = many, 5 = none)	4,4	5,0	5,0
Satisfaction	How pleasant was browsing with the current design?	4,8	3,2	2,8

*Score (1 = bad, 5 = good)

We will use these results for the next iteration of our tool. The current results can also give a preliminary conclusion on the usability of the tool. On average every user listed about 3-4 usability problems but had many suggestions. We find that in our test group, users find the tool more pleasant to work with than the leading similar websites.

8 Conclusion

In our research we suspected that user generated content can play a great role in modern search systems and similar website tools. By analyzing several search systems we found that there are many similar website search engines on the internet but no variation that uses social tagging exists yet. We focused on how the tool should function and how the usability should be implemented. We found out through testing the tags of several URLs, that when including all the tags within the search creates the most accurate results but gives back slower performance. The 10 top tags are sufficient enough to store in a database instead of all the tags for better performance. It depends on the situation which method would be best. In our case the usage would be lightweight and including all the tags works best for us. For heavier uses, we assume that a search using the top 10 or more tags will give much better performance and sufficient accuracy.

By putting ourselves in the user's perspective we designed our tool with many usability guidelines in mind. The most notable design choices for our tool is that it can adapt to many screen sizes using a fluid design, has a grid alignment of the results, and the returned list is compact with a lot of available information presented in a small area such as, website preview, top tags, tag count and user reviews.

With our tool we can prove that user generated content can indeed return accurate similar URLs. Our simple and basic evaluation of our own tool and the two leading similar website tools have shown that our tool does what it should do, find similar websites, and that the usability, especially the design, is pleasant to work with.

9 Discussion

By building this tool we address the topic of social tagging and how they can give relevant results for website similarity tools. We put a lot of effort into the project but the duration of this study didn't allow enough time for a thorough analytical research while also building the tool with all the functions we wanted. At the end, we have been able to show that our tool satisfies the need and purpose of why it's build. This study did not address the issue of how to naturally obtain URLs, users and user tags. The delicious folksonomy was also outdated, so many websites were not yet included. In the future we could expand the user involvement of the tool, where users can add their own URLs, tags instead of the URLs and tags that were obtained from the folksonomy.

With the growth of many social networks and the need for newer and better websites, this tool can satisfy both ends for similar website tools. Similar search engines should therefore incorporate user generated content for more social results and adding other social features such as reviews for a richer user experience.

10 References

Efstathiou, C.E., Central Limit Theorem. Available from: http://www.chem.uoa.gr/applets/appletcentrallimit/appl_centrallimit2.html

Devault, G., 2012. User generated content is like gold. Available from: http://marketresearch.about.com/od/market.research.social.media/a/User-Generated-Content-Is-Like-Gold.htm

Nielsen, J., 2003. Usability 101: Introduction to Usability. Available from: http://www.useit.com/alertbox/20030825.html

Nielsen, J., 2012. How many test users in a Usability Study?. Available from: http://www.useit.com/alertbox/number-of-test-users.html

O'Reilly, T., 2005. What is web 2.0. Design patterns and business models for next generation of software. Available from: http://oreilly.com/web2/archive/what-is-web-20.html

U.S. dept. of health and human services, 2006. The research-based web design & usability guidelines. Available from: http://www.usability.gov/guidelines/

Brin S. & Page, L., 1998. The anatomy of a large scale hypertextual search engine. Available from: http://ilpubs.stanford.edu:8090/361/1/1998-8.pdf

Morrison, P., 2008.Tagging and searching: Search retrieval effectiveness of folksonomies on the World Wide Web. Information Processing Management, No. 44. Available from: http://cmapspublic3.ihmc.us/rid%3D1228288503272_422770423_16390/Tagging%2520and%2520Searching.pdf

Gomez Inc. When seconds count. Available from: http://www.gomez.com/wp-content/downloads/GomezWebSpeedSurvey.pdf

English, J., Hearst, M., Sinha, R., Swearington, K. & Yee, P., 2010. Examining the Usability of Web Site Search. Berkely. Available fromhttp://flamenco.berkeley.edu/papers/epicurious-study.pdf

Heymann, P., Koutrika, G. & Garcia-Molina, H., 2007. Can Social Bookmarking Improve Web Search?Infolab Technical Report, No. 33. Available from: http://ilpubs.stanford.edu:8090/817/1/2007-33.pdf

Avrachenkov, K. & Litvak, N., 2004. Decomposition of the Google PageRankand Optimal Linking Strategy. Available from: http://hal.inria.fr/docs/00/07/14/82/PDF/RR-5101.pdf

11 Appendix

Appendix A: Analysis of different types of search systems

Search engines and methods come in different packages. From a simple search on an online store to search engines with massive databases, search engines give users the information they need and as quick as possible.

Community/Blog Search (Tested tumblr.com, blogspot.com)
These search engines uses the content of their community content to give back the data from their website. User generated content such as blog or forums posts are matched and returned to the user.

Social bookmarking websites(Tested delicious.com, pinterest.com, stumbleupon.com)
On these websites the emphasis lies with the tagging of user content. You can bookmark a link and add tags and save it for later. Instead of saving them to your favorite websites list, you are saving them to the web. Also, since the bookmarks are online, you can easily share them with others. This allows users to discover new interesting internet content.

Web directories (Tested dmoz.com)
A Web directory organizes Web sites by category. It can be compared to the Yellow Pages but on the internet which contains lists of websites specific to each category. The collections of links are usually much smaller than the other bookmarking/search engines, since the sites are manually checked by people instead of automated crawlers.

Search engines (Tested google.com, yahoo.com)
These web pages allow users to make search queries to find what they are looking for. They have huge databases containing content all over the web that were acquired automatically by the usage of spiders. The search is based on location and frequency of the search phrase within a document.

Similar website search (Tested siteslike.com, similarsites.com, similicio.us, similicio.us, similarsitesearch.com)
These websites are similar to search engines and gives back a list of usually 'popular' websites that are similar to the URL that has been queried. Users can pick a site and review it. Similar to a web directory, the site compares with other websites sharing the same category and displays them back to the user.

What is lacking?
People have shown a growing interest in community content as evidently by the increase in social media websites such as Blogs/Communities and Social Bookmarking websites. Traditional methods of finding content are search engines finding static pages or web directories showing links that are usually added or requested to be placed by the owners of the website themselves. These are both very static in nature since the user cannot contribute to the content. Even the current similar website search engines that we have tested are nothing but web directories giving back the websites that fall in the same category. Their search is thus also static web 1.0 and based on a few manual tags. People can only retrieve data and not manipulate it. Social bookmarking websites have a search tool, but are no search engines. They

also do not show websites but individual images or articles. A person looking for a few good websites that match the websites he already knows will not easily find them on these websites. A similar website search tool should be build based on social tagging to give more accurate results.

Appendix B: User scenarios

1. Simple average user

John is an average user who uses the web in his spare time. He discovered the internet and knows the sites that have been recommended to him by his friends and family. His knowledge of what is on the internet is still limited to a few popular social websites and websites to his real life interests such as football. He knows of Google search and thinks it's sufficient for every search he looks for.

Problem & Solution

A site he used to like has shutdown or doesn't satisfy his needs anymore. He needs a new website to replace this web site. He is interested in what others recommend him but cannot find this on Google. He uses the tool and fills in the URL of his 'old' favorite website and sees a list of new websites. He sorts the websites and sees which websites are the most popular and chooses to view that site.

2. Professional user

Lisa works at a web design firm as a concept designer. She is pretty good at what she does, but always looks for new places to find inspiration. She gets inspiration mostly from the work of colleagues at her own firm and others. She follows blogs and every interesting website she stumbles upon.

Problem & solution

Lisa gets an assignment from her employer to design a website for a new online Italian shoes shop. The owners want to bring something original and something that will attract young fashion people. She needs to study the competition and see how the trend is in the shoe branch. She needs inspiration so she goes on Google and can only find the more established brands and shops. She uses the similar website tool and gets an overview with thumbnails of the front pages of related/similar sites. She quickly browses through them, only opening the ones that the thumbnail appealed to her.

3. Review user

Gerard is an older businessman who uses the internet for official or business purposes. Blogs and social websites do not interest him.

Problem & Solution

He has heard that he can invest in stocks online. He wants to find websites so he can compare which online stock broker is the best. He uses Google and types in "invest stocks online". The non-sponsored results are not all related and the sponsored results are too many to find out which ones are best. He needs assurance and uses the similar website tool. He gets a quick overview of all the websites that match his keywords. He sees which websites are most tagged. But this is not enough to win him over. He uses the review button and sees the reviews of other people. He determines which website he will choose based on the negative or positive reviews.