

Synergy Between Automatic Content Generation and Social Media

Francisco Iacobelli

f-iacobelli@u.northwestern.edu

Kristian Hammond

hammond@cs.northwestern.edu

Larry Birnbaum

birnbaum@cs.northwestern.edu

Infolab
Northwestern University
2133 Sheridan Rd.
Evanston, IL 60208, U.S.

ABSTRACT

Finding out about a topic online involves visiting multiple news sites, encyclopedia entries, video repositories and other resources while discarding irrelevant information. MakeMyPage aims to speed the search process by combining automatic aggregation of information with social media to build persistent web pages with images, videos and links to important information about popular topics. Automatic aggregation provides the initial content of the web pages organized by type: blogs, news, web links, images, video and a main article. Social media provides adequate ranking of this content. MakeMyPage creates a main web page about the topic by selecting a few items from each category, and creates secondary webpages with more resources for each of these categories. Users can vote on the resources they like best and, based on these votes, links are promoted to and within the main web page in the appropriate category. In this paper, we argue that this combination of automatic retrieval and social media results in more relevant content about popular topics when compared to both traditional social media aggregators and automatic content aggregators designed to retrieve highly diversified information.

Categories and Subject Descriptors

H.3 [Information Storage and Retrieval]: Information Search and Retrieval—*Information Filtering*; H.4 [Information Systems Applications]: Information Interfaces and Presentation—*Hypertext/Hypermedia*

Keywords

Aggregation, Intelligent Information Retrieval

1. MOTIVATION

When people want to find an answer to a precise question, their first stop is usually a search engine. It is easy to query a search engine and obtain an answer. Questions such as “what are the symptoms of a common cold” or “who is Fidel Castro” can be typed verbatim into a search engine and the first results are likely to point to websites that provide these facts, such as WebMd or Wikipedia.

However, when people want to find more general information about a topic, they may have many questions that are not necessarily well-formed or relevant. After all, they are looking for a general topic because they know little about it and they may not even know the relevant questions to ask.

For example, imagine a parent who is undecided as to whether or not to home-school their children. These parents may want to find out everything about “home schooling” as a topic: the facts, best practices, relevant news about test scores of home schooled children, points of view of different people in the form of blogs, maybe some instructional videos, etc. However, when they enter “home schooling” in a search engine, they may only see a general site on how to get started, or a set of blogs that expose only one view on the matter.¹ Examples like these can be found in a variety of settings: a college student working on a paper on some political figure that he knows little or nothing about, a journalist working on an article in a topic that is new to her, a new graduate learning about a potential employer.

The people in these examples may spend considerable time digging for information on a particular subject because each website they visit provides only isolated pieces of information about the topics; furthermore, some of the information they receive may be inaccurate or irrelevant. In sum, finding out about a topic can be time consuming and by no means trivial.

These problems are exacerbated when popular topics are too recent, due to the fact that not a lot of angles on the topic have been explored and linked to yet and, therefore, search engines have not indexed enough items to provide a satisfactory level of diversity.

Now, imagine a different scenario where the people above type their topic of interest in their search engine of choice and the topmost link in the results points to a web page that is devoted to that topic and displays the most relevant resources (web links, blogs, news, facts, videos and images) about it. Such a web page would have saved these people a lot of time. Such a broad spectrum of relevant information, allows them to decide quickly what subtopics are worth focusing on, which links to explore further and which links to discard.

MakeMyPage[13] is an application that tracks popular topics and combines social media with automatic aggrega-

¹At the time of this writing, such was the first result in Google for “home schooling”

tion to produce persistent web pages that have both variety of media and good quality content about those topics. For example: if a popular search topic is Northwestern University, MakeMyPage will look for Wikipedia articles referring to Northwestern University, as well as news, blogs images and videos that are related to the institution.

In this way, the system will retrieve links to contents that are on point with the general topic: Northwestern University. However, it should be up to the community of readers, at large, to judge what contents related to Northwestern University should become more salient.

To further qualify the content retrieved based on the interests of the internet community, MakeMyPage allows users to vote links and, in this way, influence the system’s rankings. Along these lines, MakeMyPage also offers the possibility of letting users upload content, thus including new material that may be of interest to the community.

By tracking popular topics, MakeMyPage makes sure that the content it qualifies is of interest to the internet community. Because Make My Page produces persistent web pages—not just search results—that can be indexed, they will show up in the user’s favorite search engine, thus freeing the user of having to look for the same search query in multiple specialized search engines. Lastly, because people vote links with respect to the topics that generated them, the individual webpages stay relevant with respect to that topic. In particular, the best content is promoted to the top page of the set, making the most relevant information immediately salient.

Although there are content aggregators that incorporate some of these techniques, none combines automatic generation of content about intrinsically interesting topics with social media.

In this paper, we briefly present MakeMyPage and we argue that the combination of automatically generating content and social media results webpages with the most relevant content about popular topics. To show this we present two pilot user studies comparing MakeMyPage to social media websites and to leading aggregation websites.

2. THE MAKEMYPAGE PROTOTYPE

MakeMyPage web pages are comprised of a main page divided into six categories: blogs, news, web, video, article and images. For each category there is a link that says “show all” that takes the user to more resources within that category. Figure 2 shows the overall layout of a MakeMyPage. By visiting the “show all” links, users can find additional resources to vote on each section.

3. ARCHITECTURE

MakeMyPage is comprised of several modules that work together to retrieve, process, aggregate and present information. Figure 1 shows its architecture and modules. The modules from the figure are:

1. **Topic Gathering:** This module collects popular web searches from Google Trends and Tweeter. If some web searches are too close semantically, for example: “barack obama” and “president obama,” this module unifies them into one search term to avoid performing redundant searches. Semantic similarity is determined by looking at the overlap on the results of Google for each query.

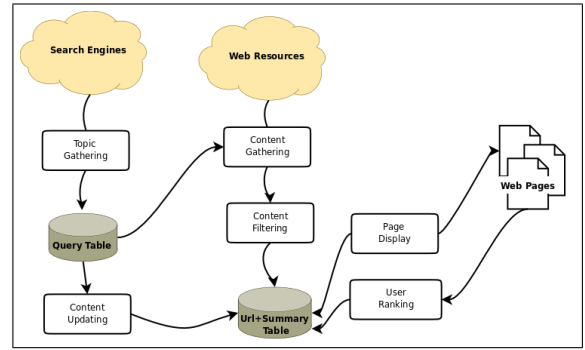


Figure 1: Overall architecture of MakeMyPage

2. **Content Gathering:** The Content Gathering module consists of a set of small modules that retrieve specialized information. Each module specializes in retrieving information in one of the following categories: web links, blogs, video, images, news and one encyclopedic article that is relevant to the search query.
3. **Content Filtering:** This module examines the data retrieved by the specialized strategies and determines which data to retain and which to throw away. Data retained is stored in the URLs database.
4. **Page Display:** This module acts as a traffic cop among modules, coordinating the resources to build and display web pages.
5. **User Ranking:** The User Ranking module controls the voting by users and the policies that regulate how to leverage the votes in order to display the items in the appropriate order.
6. **Content Updating:** This module decides whether to build new web pages or update information on existing web pages based on predefined policies.

The following section describes the techniques used to retrieve and aggregate information. These descriptions are motivated by discussion. For a general view on how they work please see [13].

4. RETRIEVING INFORMATION

Because recent popular searches are an indicators of what people consider interesting at the time, the first step to building a MakeMyPage is to extract popular and recent search terms from search engines.

The initial retrieval of content that is relevant to the search terms is supported by the Google APIs, Wikipedia’s API’s and custom APIs used to scrape popular news websites and The Internet Movie Database, IMDB.

The initial set of topics is retrieved from Google Trends – topics that have a surge in popularity within the hour. These terms are used to form queries that retrieve initial content for MakeMyPage. After this initial information is retrieved, Web links, videos and images are stored immediately in the database. The initial snippets of web links are used to retrieve meaningful keywords and entities that co-occur with the popular search terms.

These additional keywords and entities are used to form more specific and directed queries to retrieve news items,

Make My Page
TEEN CHOICE AWARDS

web

1 [2009 Teen Choice Awards, Nominees, Winners, TV Schedule](#)
Another exciting **Teen Choice Awards** presentation airs this year on FOX at 8PM ET on Monday, August 10, 2009 hosted by none other than major teen heartthrobs, ...

1 ['Twilight' Tops Teen Choice Awards Noms](#)
Jun 15, 2009 ...Wildly popular vampire romance hit 'Twilight' has nabbed the most nominations for this year's **Teen Choice Awards**.

0 [Reality TV Tryouts: Teen Choice Awards - LIVE: The TV Blog](#)
Jul 28, 2009 ... **'Teen Choice Awards'** it isn't a reality show, but it is your last chance to make your voice heard -- if you're between the ages of 13 and 19 ...

[show all web](#)

news

1 [Robert Pattinson and Kristen Stewart Attend 2009 Teen Choice Awards](#)
by CBSNews.com
This year's **Teen Choice Awards**, which is the 11th, has been kicked off and the likes of Robert Pattinson and Kristen Stewart are some of the celebs who come to the event early. Beside them, there are also Miley Cyrus, Jonas Brothers, and Zac Efron, all of them are spotted striking poses on the green grass carpet of the star-studded gala, which is staged at the Gibson Amphitheatre on Sunday afternoon, August 9 in Universal City, Calif. Kevin Jonas, Joe Jonas, and Nick Jonas manage to pose together.

0 [Obama meeting...Hudson crash...Teen Choice Awards | kxnet.com ...](#)
by KXNC
UNIVERSAL CITY, Calif. (AP) "Gossip Girl" and Zac Efron are proving their popularity among teens. The CW series and "High School Musical" star have picked up five surfboard-shaped trophies between them at the **Teen Choice Awards**.

0 [Kathy Griffin's Teen Date: Bristol Palin's Baby Daddy!](#)
by 93.9radio
Kathy Griffin definitely gets the award for most genius date at this year's **Teen Choice Awards**. Johnston wasn't only in town for **Teen Choice**. He was also here for a Vanity Fair photo shoot.

[show all news](#)

blogs

0 [Full Winner List of 2009 Teen Choice Awards in TV](#)
by eashshowbiz.com
Each collecting four kudos, 'Gossip Girls' and 'Hannah Montana' share dominance in the small screen category at this year's **Teen Choice Awards**.

article

Teen Choice Awards (Wikipedia)
The **Teen Choice Awards** is an **awards** show presented annually by FOX. The program honors the year's biggest achievements in music, movies, sports, television, fashion and more, as voted on by teens aged 13-19. The program usually features a high number of celebrities and musical performers.

video

1



Teen Choice Awards 2009 Show & Backstage Twilight Cast

0



David Beronaz & Emily Deschanel Teen Choice Awards

[show all video](#)

images

0


0


0


0


[show all images](#)

Figure 2: MakeMyPage's main page. Sections are: web links, news, blogs (not displayed), one article about the topic, videos and images.

blogs and the main article of the MakeMyPage at hand. In order to maximize the likelihood of getting relevant results, the system attempts various query strategies using entities, keywords, and combinations of both. For example, if the system does not find a wikipedia article with the terms from Google Trends, it will try to use only the main additional entity found. There are particular strategies for each category of results. These query strategies are largely based on previous research [8, 12, 18].

News, blogs and the main article go through additional filtering before being stored in the database.

4.1 News Articles

News items must be authoritative and complete. It follows then, that news items should be selected from traditionally trusted sources such as major international news agencies and newspapers. Therefore, MakeMyPage submits each search query to top newspapers and news agencies such as NY Times, USA Today, APA, Reuters, CNN, Washington Post, etc. In addition, in order to expand the sources covered, MakeMyPage submits the search terms to Google news to retrieve content from sources which may provide specialized information. For example, a news about oil prices may be well covered by a couple of traditional news sources, however Google may return related news covered by Forbes, Bloomberg and other media specialized in busi-

ness news that may provide unique perspectives or analysis about that story. In the same way, a news about crime near Sacramento, California may be covered by media giants, but a search on Google may retrieve news from local sources in Sacramento. When all these results come in (traditional and Google News), MakeMyPage visits each one and extracts the core content section of the articles. The core content section is defined as the <DIV> tag with more text in the DOM tree of the webpage's HTML source. Often times, this core content starts with a dateline; that is, the name of a place, an author, a news agency or a disclaimer that provides little or no information. Our system detects these cases using heuristics [13] and leaves only paragraphs containing some narrative. MakeMyPage, then, displays the first 3 sentences of the first paragraph (the lead) along with the title of the news story. Because this text is comprised of full sentences, the summaries that MakeMyPage displays can be read as a coherent unit of information. Moreover, because the lead and the title are usually a good abstract of the news [9, 5], the summaries of news tend to be very informative.

4.2 Blogs

Blogs reflect the opinions of users about a topic. MakeMyPage retrieves blog links, using the Google APIs, and scrapes those links to extract the section with the most text in them (see 4.1). Because there is a probability that the

blog is spam, MakeMyPage also checks to see if the contents of the blog are written in a more or less narrative way. To determine which texts were written in a narrative way we looked at the ratio of stop-words per word on the texts ($\frac{N_{stop-words}}{N_{all-words}}$). The intuition behind this metric is that often spam blogs (splogs) contain a set of keywords about many topics in order that the search engines will find them regardless of the query. Usually these keywords are not stopwords. In addition, too many stop words may result in uninteresting content. We pilot tested this intuition and, in practice, on many blogs, this ratio was a differentiating factor between spam blog and real content and the range that result in true blogs and not spam were scores between 0.4 - 0.76.

4.3 Main Article

The main article usually provides encyclopedic information about a topic or about an entity that is closely related to the topic. For example, if the search is about an important person, e.g. *Mahmoud Ahmadinejad*, Iran's president, MakeMyPage's main article should display biographical information about him. Alternatively, if the search is about an event, for example: "*Michael Jackson Toxicology*" referring to the autopsy performed on the American singer, the main article should display information about the event, such as an article on the "*Death of Michael Jackson*"; or about the important people involved, such as *Michael Jackson*; or places related to that event such as "*Neverland*".

To retrieve the main article, MakeMyPage performs a series of successive queries that stop when one query returns a result. Then that result is stored. If no results are found, no results are stored for the main article. The order of queries is as follows:

- MakeMyPage starts searching Google for Wikipedia articles using the current search term corresponding to the specific "hot trend." If a link is retrieved, then it is visited and the text of the first paragraph is stored.
- If Google does not return a result, then MakeMyPage proceeds with a second query. MakeMyPage will query the Wikipedia API for the first search term of the original query (found by the Topic Gathering module). In practice, that tends to be the main entity that users are interested in.
- If there are no articles to be found, MakeMyPage tries to search Wikipedia for relevant entities derived from the results obtained so far on the other sections. Thus, MakeMyPage concatenates the text corresponding to the four top web results and the top 2 news articles on the search terms, and sends it to OpenCalais², an entity extraction web service from Thompson Reuters. OpenCalais returns the entities detected along with a score of relevance. MakeMyPage picks the highest ranked entity and searches it using the Wikipedia API.
- If there are still no results, the main search term is searched in one last place: The Internet Movie Database (IMDB). Because the initial focus of MakeMyPage is to be good at retrieving articles about popular searches and many of these tend to be about people and popular culture, it searches IMDB, which is a good place to

find information about popular actors, TV series and movies. Therefore, if searches on Wikipedia fail, MakeMyPage searches IMDB. If IMDB returns the exact search term as one of its results, MakeMyPage visits the page for that result and stores the first paragraph of the BIO section, which contains biographical information about the person or movie title found.

5. ENHANCING RELEVANCE BY VOTING AND AUTOMATIC UPDATE OF LINKS

Despite the simplicity of the classification and filtering algorithms of MakeMyPage, the relevance of the links retrieved is good, but often times their ranking is not ideal. To improve the quality of the links that are displayed on the main page, MakeMyPage allows users to vote on content. The voting is leveraged to decide what to display on the main page. Users can vote articles up or down and each user can vote on a given article from a given web page only once. Because each link is associated to only one MakeMyPage web page, the votes for an article in one web page will not affect the ranking of a link to the same URL in a different web page. For example, a MakeMyPage about Airline accidents can share several links with the page about US Airlines. Due to a recent airline accident, the page about accidents can receive many visitors and the most relevant links would be voted up. However, a user that wants to learn about US Airlines, may want to know history, names, locations, CEOs, etc. Therefore, in this scenario, it would be detrimental to promote links on the US Airlines web page because they were voted up on the accidents web page.

Conversely, a negative vote will have no effect on other web pages that refer to the same URL and the user that casted the vote will not see that link anymore.

Users are also allowed to suggest links they find relevant. These links are scraped by MakeMyPage and, if appropriate, are put on the webpage for the topic they were suggested.

However, because some content is more time sensitive, i.e. news and blogs, the links with more votes do not necessarily appear in the front page. Each piece of news and blogs is time sensitive. In particular, recency (the time elapsed since the news was published until the present moment) is one key factor that make news stories interesting [5]. One day the news can be relevant, but the next day they may not. The same is true for blogs.

Therefore, MakeMyPage relevance algorithm relies on a ranking threshold to promote links to the main page. Link ranking is computed by the following formula:

$$\frac{1}{recency} (total - votes)$$

Here, *recency* is computed in seconds and *total - votes* is the number of positive votes minus negative votes for the given link. This method is inspired by Lerman's reverse engineering on Digg's voting policies[14].

One enhancement with respect to earlier versions of MakeMyPage[13] is that to ensure relevance of content over time, the system periodically checks its web pages and schedules updates for its content based on how much it is likely to change. MakeMyPage queries Google for each topic and sorts their posting timestamps. Then, it adds the current date as a timestamp and averages the time elapsed between timestamps. This average becomes the update interval for that topic.

²<http://www.opencalais.com/>

The next section presents background and previous work in the area of search technologies, social media and aggregation.

6. BACKGROUND AND RELATED WORK

Page rank [7] and related models are popular methods used by search engines to find relevant links about a topic, specifically links to pages that other people have found worth linking to. However, because these models rely on webpages linking to other webpages, higher ranked webpages require that people take the time to create links to them. This creates a latency effect in which the information that is necessary to rank the pages takes some time to get to the search engine.

To improve ranking algorithms, some researchers have considered using geographic information of their users [4] or presenting information as a result list and as topical clusters [10]. However, a popular method to improve ranking of links involve people's feedback in the form of votes or user comments with respect to links. Agichtein, Brill and Dumais [3] found that adding users' feedback to web searches increased accuracy of the top results by 31%. Search engines such as Google and Yahoo have been implementing user feedback for their results..

Another form of social media driven aggregation are social news aggregators [14]. Here, users post links while other users vote them up or down. Searching algorithms leverage these votes to rank the relevance of documents with respect to a search term. This strategy is successful for some topics on which users are eager to vote on, but it may not work for topics in which only a few users are interested. In one study about Digg.com, a popular social news aggregator, Lerman [15] mentions that sometimes a topic would be so interesting to users, that activity would spike and in some cases, news could be posted and voted on before Google News³ was able to index it. Social news aggregators such as Digg, with over 30 million unique users per month[2]and Reddit.com emerge as the most popular news aggregation applications on the web. Although social news aggregators can sometimes produce faster rankings of relevant information, they have a few drawbacks when searches need to be on-point about a specific topic.

Along the social aggregators, there are systems that can aggregate more than "news" and produce webpages of arranged content. For example, there are systems that allow users to search about a topic online in a collaborative manner. SearchTogether [17] is one such project that leverages automatic content retrieval with social media for ranking and organization, however the webpages users generate are not static and to produce similar results again one has to perform another collaborative search. A project that does create static webpages is GroupMe! [1] which allows users to create "wiki" like pages of media about a topic. However, GroupMe! only accepts user generated content and as such suffers from two important problems described below.

Social content aggregators present two important problems for searching content that is on point with regard to a specific topic: The first problem stems from the posting and promotion strategies of those websites. In social news aggregators, users add and promote content in a global context which disassociates it from the context in which it was

posted. Because of this disassociation, one search query can potentially return several links that are not on point with the search. For example, a search on a popular actress's name in Digg and a search on her name plus the word "pictures" may return similar results. That is because people's votes ranked the pictures very high globally, therefore, the results are retrieved regardless of the context in which they were posted. Information other than pictures, however, may not make it to the first page of results in either case. Marchionini *et al.*[16] found this problem with other social media websites as well. To solve the problem created by the global promotion of links, a system could create a unique webpage devoted to the search term, in which links are voted and promoted.

The second problem is that usually the content on these sites is driven by a few top users (most active) and those users with larger social networks[15]. This, as Lerman points out, has to do with the fact that voters can check what other users have voted on. Therefore, friends are likely to check on what other friends have voted, and consequently visit those links and vote. In addition, people "friend" popular people, such as the top users. It follows from this that people with larger social networks get their favorite sites promoted faster. It is also evident that if a topic is of no interest to those few users, it can take a long time before links with good information get enough votes to make it to the front page.

Thus, one solution for eliminating the influence that social networks have in the promotion of topics would be to find popular web searches and post links about them automatically, regardless of particular users' preferences.

Aggregation of online media does exactly this: minimize the amount of search that users have to go through to find about topics [11]. The websites stored in search engine's databases are being mined by creating custom crawlers and picking the sources and refining query terms to pull out diverse information [19, for example].

Recently, a few companies have started services that aggregate content automatically from a variety of sources although their underlying algorithms are not known. They offer an alternative to search engines [6] by providing more variety of media for a given search term. However, these services generally have not indexed nor organized very recent information. In particular, when the search query is not a frequently searched term, these websites will display standard results or nothing at all.

MakeMyPage[13], in contrast to all the approaches described above, is a system that searches for good and diverse information about a popular topic and uses automatic aggregation to generate a persistent web page of resources that are on-point with regards to the search terms, and social media to improve the quality of its contents by users that vote those links. Because the contents are tightly coupled with the generated web page, links are promoted in context.

7. RELEVANCE OF CONTENT

This paper addresses two main research questions: (a) Can we automatically generate content that is on point with a recently popular topic and that is more relevant than content generated and ranked by users about the same topic? and (b) Can we use social media to rank content about popular topics so that one main page becomes more relevant than leading aggregation search engines?

³<http://news.google.com>

In order to answer the first question, we compared MakeMyPage to leading social media aggregators Digg and Reddit in terms of relevance of links, variety and overall relevance of the results (or the web pages in the case of MakeMyPage). To answer the second question we compared MakeMyPage to Yahoo Glue and Kosmix, two leading search engines that can aggregate their results in a far more diverse and domain dependent categories than MakeMyPage.

The following sections present the method, results and discussions for two pilot studies. Study 1 addresses the first question and Study 2 addresses the second.

7.1 Study 1. Automatic Generation versus pure social media.

7.1.1 Method

In order to assess the performance of MakeMyPage content generation algorithms, we compared the quality and level of variety of the content with other social news aggregators. To select which queries to look into we took the top 100 searches from Google Hot Trends on Oct, 28, 2008 at 6:30pm. We submitted these terms to several social news aggregators⁴ and stored the webpages they returned. In addition, we created MakeMyPage web sites for the same searches. Most of the social news aggregators returned no results for most search queries extracted from Google Hot Trends. The exceptions were Digg's upcoming section⁵ and Reddit. We think that this is due to the fact that Google Hot Trends correspond to extremely recent searches, therefore, users of social aggregators may take some time to discover good links and post about them. Out of the searches that produced results in Digg, Reddit and MakeMyPage, we randomly selected 5 for comparison.

To evaluate MakeMyPage against these web sites we designed a pilot study. We reformatted the Reddit and Digg pages for the five searches by stripping them of background color and logos, leaving a plain website that was somewhat faithful to the original layout, but with no identifying information. Similarly, we stripped almost all formatting from the MakeMyPage web sites for the searches, leaving the original layout. We left the layouts because they also carry information. For example, a layout that has pictures along with links carries different information than having all links first and then all pictures, disassociated to the links they were attached to originally. Figure 3 shows sample snippets of what each website looked.

For the pilot study we had 11 participants. All of them adults. They either received an email with a URL to participate in the study or they participated because someone told them about the study and provided the URL. Two of the participants were vaguely familiar with MakeMyPage. The rest were not familiar at all with MakeMyPage. We randomly assigned one of the five query terms to each participant and they saw the three web sites in random order. The participants then had to rate five items from 1 to 10, and then specify which web site helped them learn more

⁴We searched on Digg's front page, Digg's upcoming, Reddit's front page, Mixx's front page and StumbleUpon's front page.

⁵Digg's upcoming section contains brand new links that have been voted, but do not add enough votes yet to be in the front page of Digg. The most voted links usually make it to the front page in a matter of hours or minutes.

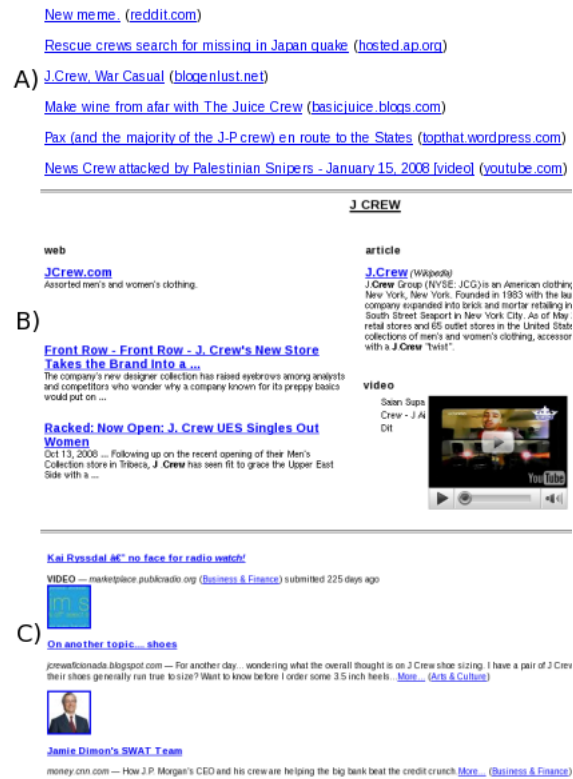


Figure 3: Websites used for the experiment. The figure shows snippets of the webpages generated for the search “J. Crew.” A) Reddit.com, B) MakeMyPage and C) Digg.

about the search query. The five items were (a) relevance of the content of the page, (b) relevance of the links in the webpage, (c) variety of relevant information about the search, (d) How much they learned about the search terms and (e) How much they knew about the query prior to taking the test.

7.1.2 Results

The results showed that, in general, users knew little about the topics. The mean rating for previous knowledge of topics in the survey was 2.5 (SD=2.6, Median=1). For the rest of the items, we analyzed the data using multiple paired t-tests to compare the three websites on each item in the survey. MakeMyPage scored significantly better than the other two websites in all items of the survey. Because in all items of the survey Digg and MakeMyPage obtained the highest scores, I will provide the results of the paired t-test comparison among those two only. For a graphic comparison of the three means, please look at Figure 4

On item a, The overall relevance of the content of the page, a paired t-test shows highly significant: $t(10)=3.96$; $p<0.01$, on item b, the relevance of the links displayed on the webpage, the t-test comes significant again: $t(10)=3.9$; $p<0.01$. Item c, the variety of information about the query referred to the number of different media to convey information about the search. On this item, MakeMyPage also showed significantly higher scores. $t(10)=4.083$; $p<0.01$. The last item to rate was how much was learned about the search

term. MakeMyPage's average score was significantly higher than Digg's. $t(10) = 6.83$; $p < 0.01$

7.1.3 Discussion

It is possible that the crowd that visits Reddit and Digg may be different than the people that are likely to go to MakeMyPage. However, because Reddit and Digg are the leading social media websites for recent and popular topics and because MakeMyPage aims to create webpages on these topics, we believe they are a fair baseline to compare a social media systems that generate content on these popular topics. Because MakeMyPage aggregates different kinds of online media, we were expecting good results on item (c), variety of information. For item (d), amount learned about the search term, we expected less of a difference between Digg and MakeMyPage. This is because it is easy to find out information by skimming text on result sets so, when the text is informative this process is faster. In particular, when Digg users post a link, they also post a little summary about the news, which, intuitively, should be more informative than just extracting a paragraph of the story, like MakeMyPage does. Despite this advantage for Digg and Reddit, MakeMyPage scored better in terms of amount learned about the search query. We think this is tightly related to the variety of information provided by MakeMyPage's webpages together with excerpts that consist of full sentences. By providing both graphic, video and textual media, users can get a good deal of information just by browsing the main page.

Another surprise was the scoring for item (b), relevant links. We did not expect such a big difference between MakeMyPage and Digg. The difference here may be driven by the fact that in a social news aggregator the content and ranking of links is driven by users with the largest social networks [15] and what we see in their results is basically links that are relevant to a few users in some dimension. Moreover, most of these links offer repeated information. If this information turns to be somewhat relevant, then it is likely that most links in the result set are equally somewhat relevant. MakeMyPage, in contrast, retrieves links from several resources and therefore, it is likely to retrieve some links that are less relevant as well as links that are significantly more relevant to average users.

However, there was one case in which links were equally relevant, yet people learned more from MakeMyPage. Users who saw a search about a popular model: Lydia Hearst, rated the relevance of links in MakeMyPage and Digg equally. Because she is such a popular celebrity there are many links to different pieces of information about her. However, while in Digg all links were along similar dimensions (photo galleries and gossip) in MakeMyPage there were at least two more dimensions: biographical links and blogs. This diversity of links was reflected by the fact that those users who saw the webpages about Lydia Hearst assigned the same score to Digg and MakeMyPage on the relevance of the links, but assigned MakeMyPage a higher score in terms of amount of information learned.

The results for Digg on Lydia Hearst illustrate one major strength of MakeMyPage: as we discussed in the background section, Digg and other social news aggregators suffer from the problem of globally promoting links regardless of the context in which they were generated [16]. In the case of Lydia Hearst, gossip and pictures were highly ranked because

users voted the picture galleries in numbers much greater than other links about her. The result is a search result on the most popular links about Hearst, and since they are not attached to a context, they turn out to be very similar. However, because our algorithms aggregate from different media, such as blogs, news and web links, MakeMyPage was less likely to repeat the same information, thus effectively providing links that were relevant to the search on more than one dimension (for example, biographies, opinions, videos, news stories, etc.). In sum, it may be the case that in Digg or Reddit, the most popular users are posting the single link that is more relevant to them, thus pushing down the ranking of links that are relevant to other, less popular, users; and MakeMyPage levels this disparity in ranking.

The results presented here correspond to a pilot study and, as such, there are many aspects that can be improved for a full study. First, the searches we used were only five. More search queries may reveal the domains for which MakeMyPage performs best and domains for which social news aggregators or search engines may be more useful. Second, the searches we used were taken from Google Hot Trends, whose nature is to have searches with a very recent and spiked surge in activity. This means that the relevant topics may not make it immediately to social news aggregators. A broader study should consider searches about topics that are older and therefore have more mature content in these social news aggregators and other social media websites. Third, a broader study should also consider more participants and more websites for comparison as well as questions about the topics to assess amount of information learned. Lastly, in future studies we have to consider some basic modifications in the methodology, such as varying the age of participants and frequency of use of social media, and asking not only for relevant links, but interesting links as well. It may be the case that Digg and Reddit users are not voting on links that are highly relevant to a topic, but that are interesting to them at that particular time.

7.2 Study 2. Social media to rank relevant content

7.2.1 Method

MakeMyPage is unique in its approach to user interaction because it is not a search engine, but a collection of persistent web pages that can be indexed by search engines. Therefore, to assess the relevance of the links automatically retrieved and our ranking mechanisms we compared the quality and level of variety of the content with two major search engines that produce aggregated content in a similar way to MakeMyPage: Yahoo Glue and Kosmix. To select which queries to compare, we took search terms, chosen at random, from the top 100 searches from Google Hot Trends on January 9th, 2008 at 11 pm. and three random terms from the top 10 Yahoo Buzz (popular searches on Yahoo) at the same time. We submitted the search terms to Yahoo Glue and Kosmix, however, not all the search terms generated results in Kosmix or Yahoo Glue, so we submitted terms randomly until we had six web pages for three terms from Google Trends and three terms from Yahoo Buzz. We also created MakeMyPage web sites for those six terms.

To evaluate MakeMyPage against these web sites we designed a pilot study. We reformatted the Yahoo Glue and Kosmix pages for the six searches by stripping them of lo-

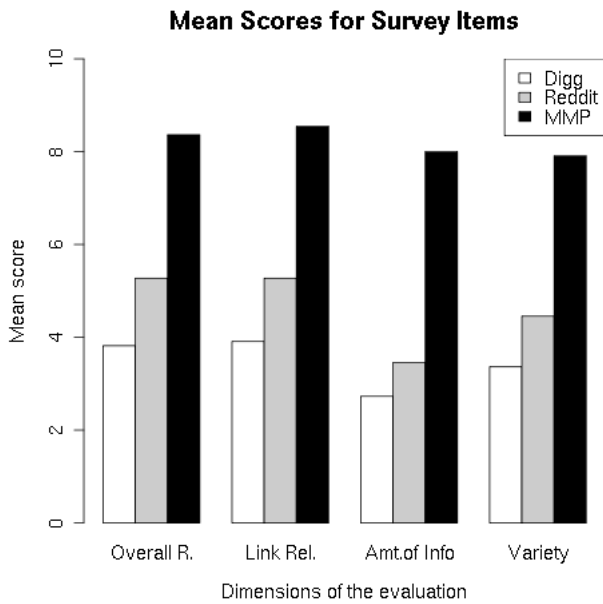


Figure 4: Comparison of the mean scores of the three sites on: overall relevance (item a), link relevance (item b), amount learned (item c) and variety (item d)

gos, proprietary links (such as “my profile”, or “FAQ”) and identifying links. This process left the websites as faithful as possible to the original layout, but with no identifying information. We processed the six MakeMyPage web sites in the same manner. We strove to leave the layouts intact because they also carry editorial information that can make a page more useful and diversity more salient. Figure 5 shows sample snippets of what each website looked like.

For the pilot study we had 43 participants, all of them adults. Most of them had taken at least one Computer Science class and all of them use search engines regularly. None of them knew about the websites in the study. They received an email with a URL to participate in the study. The study was carried out in two stages: first, 28 participants compared the three website on one of the six terms. We randomly assigned one of the six query terms to each participant and they saw the three web sites in random order. The participants then had to rate five items from 1 to 10, and then specify which web site helped them learn more about the search query. The five items were (a) relevance of the content of the page, (b) relevance of the links in the web page, (c) variety of relevant information about the search, (d) how much they learned about the search terms and (e) how much they knew about the query prior to taking the test. After they rated the websites they were asked to vote on the most relevant links for the topic on a fully functional MakeMyPage web page. This ended the first part of the study. A second group of 15 people rated the web pages using the same methodology, but this time, the MakeMyPage version of the web pages corresponded to the voted web pages for the search terms, again, stripped of any identifying information. The second group did not vote on further web pages. We then compared the ratings of all

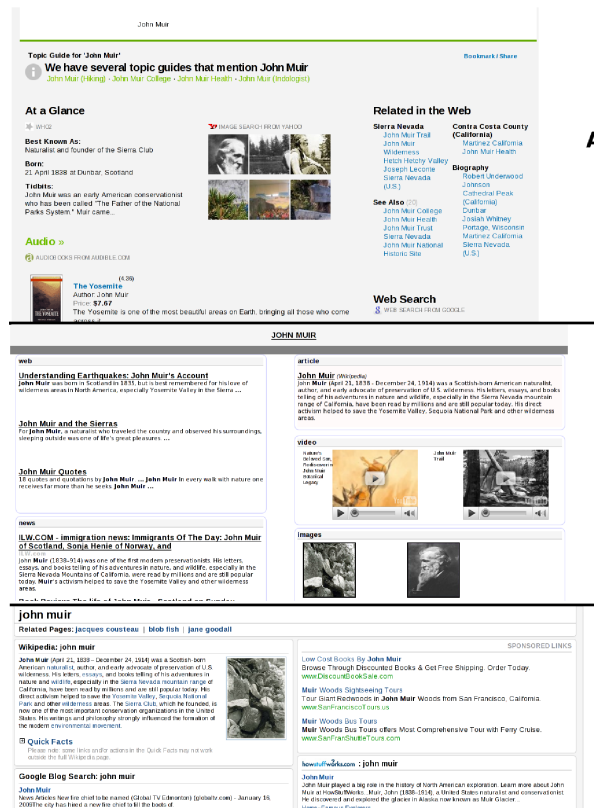


Figure 5: Websites used for the experiment. The figure shows snippets of the web pages generated for the search “John Muir” A) Kosmix, B) MakeMyPage and C) Yahoo Glue.

three before and after the voting happened.

7.3 Results

The results showed that, in general, users knew little about the topics. The mean rating for previous knowledge of topics in the survey was 3.37 ($SD = 2.4$, $Median = 2$). For the rest of the items, we analyzed the data using multiple paired t-tests to compare the three websites on each item in the survey. Before users voted on the content of MakeMyPage, it scored significantly better than Yahoo Glue in all items of the survey. However, Kosmix scored significantly higher than MakeMyPage. However, after voting on the MakeMyPage websites, the differences between Kosmix and MakeMyPage relevance scores (amount learned, overall relevance of the web page and relevance of the links within the web page) were not significant. However, the ratings for MakeMyPage relevance of links improved significantly: $t(27) = -2.4; p < 0.05$. Moreover, after voting, MakeMyPage showed a significantly higher ranking for link relevance when compared to Yahoo Glue: $t(14) = 2.64; p < 0.05$. For a graphic comparison of the three websites on this item, before and after voting happened, please look at Figure 6

7.3.1 Discussion

Kosmix and Yahoo Glue not only display videos, photos, articles, news and blogs, but they also search for prices on items related to the topic, opinions in forums, discography or books written by or about persons related to the topic and so

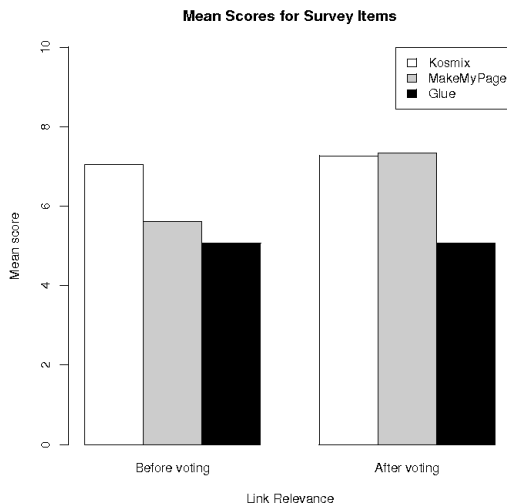


Figure 6: Comparison of the mean scores for link relevance of the three sites before and after voting on MakeMyPage

on. For this reason, we did not expect MakeMyPage to outperform them in the item related to variety of information, although Yahoo Glue was not rated significantly higher than MakeMyPage in this respect.

However, after voting, users should only bring to the front what is really useful and on-point. It is not surprising, then, that after voting, the front page of the MakeMyPage websites was rated significantly higher than Yahoo Glue and, although not significant, slightly higher than Kosmix. What this shows is that MakeMyPage web pages bring results that are comparable to those of the other two websites despite its less diverse content pool. MakeMyPage can do this by bringing results that people find truly relevant about a topic and not the mere results of a web search on the topic.

It is worth noting that although the relevance of the links of MakeMyPage were up to par with the other two aggregators, it was not easy to find search terms that generated webpages on all three websites on study 1 and 2. To measure which system provided the most webpages about popular topics, we randomly selected 20 search terms from Google Hot Trends and submitted them to MakeMyPage, Kosmix and Yahoo Glue. MakeMyPage created webpages for all 20. Kosmix created webpages for 9 of them and Yahoo Glue only for 5. This difference was statistically significant ($\chi^2(24.5, 2)$; $p < 0.001$). We think that for practical purposes, people may benefit more from a software like MakeMyPage in that the content can quickly be ranked to be as relevant as the best aggregator we compared to and users can count on a MakeMyPage webpage for almost every popular search.

The results presented here correspond to a pilot study and, as such, there are many aspects that can be improved for a full study. Increasing the number of search terms, analyze results by domain of search, mix traditionally popular search terms with more recent ones, etc. Additionally, link relevance is not the only measure for the relevance of a website. Therefore, future studies should consider more precise and detailed measures of content relevance, such as explor-

ing single “most relevant” and “least relevant” links for the query, or explore relevance by category.

8. FUTURE WORK

Because the improvements of voting on MakeMyPage over Kosmix were not statistically significant, we should consider two alternative hypothesis to guide future work: (a) user feedback (voting) can be overcome by greater diversity of content (Kosmix) or by smarter ranking algorithms (after all, the MakeMyPage webpages did contain content that was, at least, as relevant as Kosmix’s, although poorly ranked initially)

Thus, we are expanding our efforts to improve both diversity of content and ranking algorithms. We are also exploring the feasibility of task dependent intelligent aggregators. That is, an aggregator whose algorithms are fine tuned to provide relevant and on-point information for a specific task. In particular, we are exploring the feasibility of building online news reading interfaces using intelligent aggregation algorithms [12].

9. CONCLUSIONS

Make My Page aims to be a new kind of social media system that automatically generates new content collections based on popular demand and then opens the editing process up to the collective judgment of users. Tracking the queries submitted to search engines, the system compiles collections of persistent web pages, news, images and videos driven by those queries and makes them available to end users. To do this from the starting point of search, it utilizes three core technologies. First, it makes use of filtering technologies to ensure that the initial content it retrieves is relevant. Second, it uses a set of intelligent ranking techniques that we have developed to further filter and order its results before presenting them. Finally, on the presentation side, it encourages end users to provide feedback on the results in the style of Digg and other social media sites.

We argue that MakeMyPage retrieval algorithms result in more relevant content about popular topics than traditional social aggregators such as Digg and Reddit do. We also argue that social media can help rank information in such a way that the main webpage becomes as informative as the more extensive pages of leading content aggregators such as Kosmix or Yahoo Glue. Finally, because of the ability to retrieve webpages from popular searches, we argue that MakeMyPage may be a better resource of information for very recent popular queries.

We hypothesize that because the content will be on-point with regard to popular queries and because users will keep it properly ranked, the pages produced by MakeMyPage will tend to percolate to the top of the result sets. Because they contain valuable content, they will tend to stay there. This is a departure from the model of social aggregators in which users generate content and it is a departure from automatic aggregators that double as search engines.

10. REFERENCES

- [1] F. Abel, M. Frank, N. Henze, D. Krause, and P. Siehdnel. Groupme! – combining ideas of wikis, social bookmarking, and blogging. In *International Conference on Weblogs and Social Media*, March 2008.

- [2] J. Adelson. Big news: Expanding & growing digg. <http://blog.digg.com/?p=256>., September 2008. Digg Company. Retrieved on Nov 3, 2008.
- [3] E. Agichtein, E. Brill, and S. Dumais. Improving web search ranking by incorporating user behavior information. In *proceedings of ACM SIGIR '06*, pages 19–26, 2006.
- [4] R. B. Almeida and V. A. F. Almeida. A community-aware search engine. In *Proceedings of the 13th international conference on World Wide Web*, pages 413–421, 2004.
- [5] A. Bell. *The Language of News Media*. Language in Society. Wiley-Blackwell, September 1991.
- [6] P. Bradley. Search Engines: New Search Engines in 2006. *Ariadne, ISSN 1361-3200*, 2007.
- [7] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, April 1998.
- [8] J. Budzik, K. J. Hammond, and L. Birnbaum. Information access in context. *Knowledge-Based Systems*, 14(1-2):37–53, March 2001.
- [9] Cappon. *Associated Press Guide to News Writing: The Resource for Professional Journalists*. Arco, 3 edition, 1991.
- [10] P. Ferragina and A. Gulli. A personalized search engine based on web-snippet hierarchical clustering. *Softw. Pract. Exper.*, 38(2):189–225, 2008.
- [11] A. Halevy, A. Rajaraman, and J. Ordille. Data integration: the teenage years. In *Proceedings of the 32nd international conference on Very large data bases*, pages 9–16, 2006.
- [12] F. Iacobelli, L. Birnbaum, and K. J. Hammond. Tell me more, not just "more of the same". In *Intelligent User Interfaces (IUI2010)*, February 2010.
- [13] F. Iacobelli, K. Hammond, and L. Birnbaum. Makemypage: Social media meets automatic content generation. In *ICWSM 2009*, 2009.
- [14] K. Lerman. Dynamics of collaborative document rating systems. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 46–55, 2007.
- [15] K. Lerman. User participation in social media: Digg study. In *International Conference on Web Intelligence and Intelligent Agent Technology*, volume 0, pages 255–258, 2007.
- [16] G. Marchionini, R. Capra, and C. Shah. Focus on results: Personal and group information seeking over time. In *HCIR2008*. Microsoft Research, 2008.
- [17] M. R. Morris and E. Horvitz. Searchtogether: an interface for collaborative web search. In *UIST '07: Proceedings of the 20th annual ACM symposium on User interface software and technology*, pages 3–12, New York, NY, USA, 2007. ACM.
- [18] N. Nichols and K. Hammond. Machine-generated multimedia content. In *ACHI '09: Proceedings of the 2009 Second International Conferences on Advances in Computer-Human Interactions*, pages 336–341, Washington, DC, USA, 2009. IEEE Computer Society.
- [19] A. Wright. Searching the deep web. *Communications of the ACM*, 51(10):14–15, 2008.