Exploiting Social Networks for Internet Search

Alan Mislove[†‡] Krishna P. Gummadi[†] Peter Druschel[†]

[†]Max Planck Institute for Software Systems, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany.
[‡]Rice University, 6100 Main Street, MS-132, Houston, TX 77005, USA.

1 INTRODUCTION

Over the last decade, the World Wide Web and Web search engines have fundamentally transformed the way people find and share information. Recently, a new form of publishing and locating information, known as online social networking, has become very popular. While numerous studies have focussed on the hyperlinked structure of the Web and have exploited it for searching content, few studies, if any, have examined the information exchange in online social networks.

In the Web, explicit links called hyperlinks between content (typically pages) are the primary tool for structuring information. Hyperlinks are used by authors to embed a page in the Web of related information, by human users to manually browse the Web, and by search engines to crawl the Web to index content, as well as to rank or estimate the relevance of content for a search query.

In contrast to the Web, no explicit links exist between the content (typically photos, videos, and blog postings) stored in social networks. Instead, explicit links between users, who generate or publish the content, serve as the primary structuring tool. For example, in social networking sites like MySpace [15], Orkut [17], and Flickr [4], a link from user *A* to user *B* usually indicates that *A* finds the information published by *B* interesting or relevant, or *A* implicitly endorses *B*'s content due to an established social relationship. Such social links enable users to manually browse for information that is likely of interest to them, and could be used by search tools to index and locate information. In this paper, we seek to understand whether these social links can be exploited by search engines to provide better results.

This paper makes three contributions: First, we compare the mechanisms for content publication and location in the Web and online social networks. We argue that search techniques could benefit from integrating the different mechanisms used to find relevant content in the Web and social networks. Second, we present results from an experiment in social network-based Web search to support our contention. Third, we outline the research challenges and opportunities in leveraging social networks for future Internet search.

The remainder of this paper is organized as follows. In Section 2, we first contrast how content is exchanged in social networks and the Web, and then speculate on the potential of integrating the different search techniques used in these systems. We evaluate the potential of our integrated approach to search using a social network-based experiment in Section 3. We discuss the research challenges that need to be addressed in order to realize such an integrated search system in Section 4. We present related work in Section 5 and conclude in Section 6.

2 THE WEB VERSUS SOCIAL NETWORKS

In this section, we compare the Web and social networking systems, with respect to their mechanisms for *publishing* and *locating* content.[1] Publishing refers to the mechanism by which content creators make information available to other users; it includes the way users relate their content to other content found in the system. Locating refers to the mechanism by which users find information relevant to them; it includes the ways users browse or search the content in the system.

2.1 The Web

In the Web, the content typically consists of Web pages written in HTML.

Publishing: Users publish content by placing documents on a Web server. An author places hyperlinks into her page that refer to related pages. She may also ask other authors to include links to her page in their pages. Often, such links are placed deliberately to ensure the page is indexed and ranked highly by search engines.

Locating: Today, the predominant way of locating information on the Web is via a search engine. Modern Web search engines employ sophisticated information retrieval techniques and impressive systems engineering to achieve high-quality search results at massive scale.

The key idea behind search engines like Google is to exploit the hyperlink structure of the Web to determine both the corpus of information they index and the relevance of a Web page relative to a given query [18]. This

[1] We ignore the mechanisms for distributing content between users as they are similar in both the Web and many current online social networks. In both systems, the content is transferred using HTTP over TCP, and the users navigate the systems using their Web browser.

approach has proven highly effective, because the incident links to a page are strong indicators of the importance or relevance of the page's content in the eyes of other users.

However, hyperlink-based search has some well known limitations. First, while Web search is very effective for relatively static information, it may under-rate or miss recently published content. For a new page to be noticed and appropriately ranked by a search engine, (a) it must be discovered and indexed by the search engine, (b) hyperlinks to the new page must be included in subsequently published or edited pages, and (c) all such links must then be discovered by the search engine.

Second, as search engines determine the relevance of a page by its incident hyperlinks, their rating reflects the interests and biases of the Web community at large. For instance, a search for "Michael Jackson" yields mostly pages with information about the pop star. Computer scientists, however, may find the Web page of a professor with the same name more relevant. Refining the search to find that page is possible but can be tricky, particularly if one does not recall the professor's current affiliation or field of specialization.

Finally, the hyperlink structure influences whether a page is included in a search engine's index. Unlinked pages and non-publicly accessible pages are not indexed. Many other pages are not indexed because the search engine deems them insufficiently relevant, due to their location in the hyperlink structure. As a result, obscure, special-interest content is less likely to be accessible via Web search.

2.2 Social Networks

Online social networking Web sites have recently exploded in popularity. Sites offer services for finding friends like MySpace [15], Orkut [17], and Friendster [6], for sharing photos like Flickr [4], for sharing videos like YouTube [24] and Google Video [8], and for writing blogs like LiveJournal [12] and BlogSpot [3]. These sites are extremely popular with users: MySpace claims to have over 100 million users, while Flickr and Orkut boast 2.5 million and 13 million users, respectively. MySpace recently has been observed to receive more page hits than Google [16].

Examples of online social networking, though, have existed for much longer. For instance, the common practice of placing content on the Web and sending its URL to friends or colleagues is essentially an instance of social networking. Typically, the author has no intention of linking the content; thus, the content remains invisible to users other than the explicit recipients of the URL. The content is advertised not via hyperlinks, but via links between users.

Publishing: Users publish content by posting it on a social networking site. Content is associated with the user who introduced it, and with users who explicitly recommend the content. Explicit links do not generally exist between content instances, and the content can be of any type. Often, the content is temporal in nature (e.g. blog postings), non-textual (e.g. photos and video clips), and may be of interest only to a small audience.

Independent of the content, users maintain links to other users, which indicate trust or shared interest. Links can be directed (indicating that the source trusts or is interested in the content of the target) or undirected (indicating mutual trust or interest in each other's content). Some systems maintain groups of users associated with different topics or interests; users can then join groups rather than specifying links to individual users. In some systems, the full social network graph is public; in others, only immediate neighbors of a node can view that node's other neighbors.

Locating: The predominant method of finding information in online social networks is to navigate through the social network, browsing content introduced or recommended by other users. Some sites also provide keyword-based search for textual or tagged content. Additionally, other sites have 'top-10' lists showing the most popular content, where the popularity is determined according to how often users have accessed the content or based on explicit recommendations provided by users.

Moreover, social networks enable users to find timely, relevant and reliable information. This is because users can browse adjacent regions of their social network, which likely consist of users with shared interests or mutual trust. Since the content can be non-textual, obscure, or short-lived, it may be hard to find by the way of Web search. For example, blog posts are generally of short-term interest, videos and photos are non-textual, and all three types of content tend to be of interest to a limited audience.

Content in social networks can also be rated rapidly, based on implicit and explicit feedback of a large community of content consumers. In contrast, Web search relies on the slower process of discovering hyperlinks in the Web, which are created by a relatively smaller number of content authors. Since content rating in social networks is performed by the content consumers, rather than the producers, content introduced into the network can by rated almost immediately.

2.3 Integrating Web search and social networks

Today, the information stored in different social networks and in the Web is mostly disjoint. Each system has its own method of searching information. While search companies have started to address this issue with specialized search tools for RSS-based news feeds and for

blogs, there is no unified search tool that locates information across different systems. Social network-based search methods are not generally used in the Web, though services like Google Scholar support search facilities tailored to a specific community. Given that end users access both the Web and the social networks from the same browsers, it seems natural to unify the methods to find information as well.

In this paper, we explore the idea of integrating Web search with search in social networks. We believe that such an approach could combine the strengths of both types of systems: simultaneously exploiting the information contained in hyperlinks, and information from implicit and explicit user feedback; leveraging the huge investment in conventional Web search, while also ranking search results relative to the interests of a social network; and locating timely, short-lived, non-textual or special-interest information alongside the vast amounts of long-lived and textual information on the Web.

3 PEERSPECTIVE: SOCIAL NETWORK-BASED WEB SEARCH

Our discussion above suggests that (a) a growing body of Internet content cannot be retrieved by traditional Web search as it is not well-connected to the hyperlinked Web, and that (b) social network links can be leveraged to improve the quality of search results. We are currently exploring the benefits of social networks-based Web search as part of the *PeerSpective* project. In this section, we describe a simple experiment we conducted to validate and quantify our two separate hypotheses.

3.1 Experimental methodology

We recruited a group of ten graduate students and researchers in our institute to share all Web content downloaded or viewed with one another. Each user runs a lightweight HTTP proxy, which transparently indexes all visited URLs. When a Google search is performed, the proxy transparently forwards the query to both Google as well as the peer proxies of other users in the social network. Each proxy executes the query on the local index and returns the result to the sender. The results are then collated and presented alongside the Google results as shown in Figure 1.

Our experimental PeerSpective prototype relies on the Lucene [13] text search engine and the FreePastry [5] peer-to-peer overlay. We configured Lucene to follow Google's query language. Also, we ranked the results obtained from PeerSpective by multiplying the Lucene score of a search result by the Google PageRank of that result and adding the scores from all users who previously viewed the result. Thus, PeerSpective's ranking takes advantage of both the hyperlinks of the Web (via Google's PageRank) and the social links of the user community.

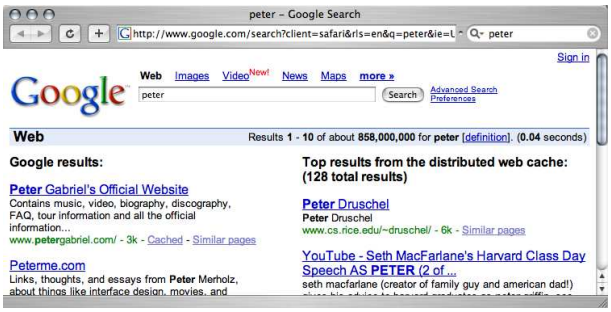

Figure 1: **Screenshot of our PeerSpective search interface**. Results from the distributed cache appear alongside the normal Google results.

We present measurements and experiences from a one month long experimental deployment. During this time, the 10 users issued 439,384 HTTP requests covering 198,492 distinct URLs. Only 25.9% of the HTTP requests were of content type `text/html` or `application/pdf`, meaning they could be indexed by our proxy. The remaining requests consisted of images, javascript, and other miscellaneous types.

Given that our user base is small, includes the authors, and represents a single community with highly specialized interests, we cannot claim that our results would be representative of a deployment with a larger, diverse user base. However, we believe our results indicate the potential of social network-based Web search. A more comprehensive study, which also considers Web access traces collected at the gateway router of a major university, is currently in progress.

3.2 Limits of hyperlink-based search

Even the best Web search engines do not index content that is not well linked to the general Web or content that is not publicly available. So, our first goal is to understand and quantify the Internet content that is viewed by users, but is not captured by the search engines. We would also like to know how much of this content is already indexed by another user in PeerSpective.

To estimate the limits of hyperlink-based search, we check what fraction of the URLs actually visited by the users are not indexed by Google. There are a number of reasons why a page may not be indexed by Google: (a) the page could be *too new*, such a as blog posting or news article; (b) the page could be in the *deep web* and not well-connected enough for Google to choose to crawl it; or (c) the page could be in the *dark web*, where it is not publicly available or is not referred to by any other page.

For each HTTP request, we checked whether Google's index contains the URL, and if some peer in PeerSpective has previously viewed the URL. Since search engines

URL	Too new	Deep web	Dark web
http://jwz.livejournal.com/413222.html	✓	✓	
http://www.mpi-sws.mpg.de/∼pkouznet/ ... /pres0031.html		✓	
http://sandiego.craigslist.org/w4m/179184549.html	✓	✓	
http://edition.cnn.com/2006/ ... /italy.nesta/index.html	✓		
http://72. ... .163/status.asp			✓
http://www.itv.com/news/ ... a8e4b6ea.html	✓		
http://www.stat.rice.edu/∼riedi/ ... /target21.html		✓	
http://amarok.kde.org/forum/index.php/board,9.20.html	✓	✓	

Figure 2: **Sample URLs that were not indexed by Google**. We manually inspected the URLS to determine the likely reason for not being in Google's index, as discussed in Section 3.2.

only index static HTML content, we considered only URLs of indexable content types which did not have any GET or POST parameters and ended in either `.html` or `.htm`. Further, we discarded URLs with an auto-refresh feature (such as the scoreboard sites for sports), as they would artificially bias the results against Google. This left us with 6,679 requests for 3,987 URLs.

Our analysis shows that Google's index covers only 62.5% of the requests, representing 68.1% of the distinct URLs. This implies that about one third of all URLs requested by our users cannot be retrieved by searching Google! Our analysis also showed that the union of the PeerSpective peer indexes covers about 30.4% of the requested URLs. While PeerSpective achieves only half of the coverage of Google's index, it does this with a much smaller size: at the end of the experiment, the PeerSpective indexes contained 51,410 URLs, compared to Google's index of over 8 billion URLs.

Additionally, we found that 13.3% of the URLs viewed were contained in PeerSpective but not in Google's index. These documents were not available via Google's search engine but had been requested before by someone in the peer network. This increase in coverage amounts to a 19.5% improvement by PeerSpective compared to normal Google search. It is worth noting that, for our small social network of computer science researchers, this improvement in coverage was possible by adding just a few thousand URLs to a Google index containing billions or URLs.

Our results naturally raise the question, what are these documents that are of a of interest to our users, but are not indexed by Google? We manually analyzed a number of such URLs and show a random sample of them in Figure 2. We additionally list the likely reasons why each URL does not appear in Google's index.

3.3 Benefits of social network-based search

Another challenge facing search engines is ranking all the indexed documents in the order of their relevance to a user's query. Ranking is crucial for search, as most users rarely go beyond the first few query results [20]. Our goal here is to study how often users click on query results from PeerSpective as opposed to Google. As shown in Figure 1, our users are presented with results from both Google and PeerSpective for every Google query.

During the course of the month, we observed 1,730 Google searches. While Google's first result page contained an average of 9.45 results, our smaller PeerSpective index resulted in an average of 5.17 results on the first page. Of the 1,730 queries, 1,079 (62.3%) resulted in clicks on one or more search result links, 307 (17.7%) were followed by a refined query, and after the remaining 344 (19.8%), the user gave up. We found that 933 (86.5%) of the clicked results were returned only by Google, 83 (7.7%) of the clicked results were returned only by PeerSpective, and 63 (5.7%) of the clicked results were returned by both. This amounts to a 9% improvement in search result clicks over Google alone, as 83 of the search result clicks would not have been possible without PeerSpective.

It should be kept in mind that this 9% improvement over Google, considered by many to be the gold standard for Web search engineering, was achieved by a simple, very small, social network-based system quickly put together by three systems researchers over a period of a few days. Based on our early experience, we feel that these results suggest inherent advantages of using social links for search, which could be exploited better with more careful engineering.

3.4 Discussion

To better understand the cases when PeerSpective search results outperform Google results, we manually analyzed the corresponding queries and result clicks. We show a random sample of the data we analyzed in Figure 3. We observed that the reasons for clicks on PeerSpective results fall into three categories:

Disambiguation: Some search terms have multiple meanings depending on the context. Search engines generally assume the most popular term definition. Social networks can take advantage of the fact that communities tend to share definitions or interpretation of such terms. An example for disambiguation is shown in Figure 3, where a user's query for "bus" yielded the local

Query	Page clicked on	Disambiguation	Ranking	Serendipity
bus	Saarbrücken bus schedule	✓	✓	
stefan	FIFA World Cup site			✓
peter	Peter Druschel's home page	✓		
serbian currency	XE.com exchange rates		✓	
coolstreaming	CoolStreaming INFOCOM paper		✓	
moose	Northwest Airlines' contract of carriage			✓
münchen	Peter Druschel's homepage			✓

Figure 3: **Sample search queries for which PeerSpective returned results not in Google**. The results are categorized into different scenarios discussed in Section 3.4.

bus schedule, as it is the page with this keyword that is most visited by local users in the network.

Ranking: Search engines rank all relevant documents and return the top of the resulting list. Social networks can inform and bias the ranking algorithm, since nearby users in the network often find similar sets of pages relevant. An example we observed is a search with the term "coolstreaming". A Google search ranks most highly popular sites (such as Wikipedia) discussing the CoolStreaming technique for P2P streaming of multimedia content. PeerSpective ranked the INFOCOM paper describing CoolStreaming at the top, as it is most relevant to our researchers.

Serendipity: While browsing the Web, users often discover interesting information by accident, clicking on links that they had not intended to query for. This process, termed serendipity, is an integral part of the Web browsing experience. Search results from PeerSpective provide ample opportunity for such discoveries. For example, while looking for information about "München" (Munich), one of our users discovered that a fellow researcher attended school in München, thus finding a convenient source of information about the city.

4 OPPORTUNITIES AND CHALLENGES

Online social networking enables new forms of information exchange in the Internet. First, end users can very easily and conveniently publish information, without necessarily linking it to the wider Web. Second, social networks make it possible to locate and access information that was previously exchanged by "word of mouth", that is, by explicit communication between individuals. Third, unlike Web search engines, which organize the world of information according to popular opinion, social networks can organize the world of information according to the tastes and preferences of smaller groups of individuals.

We see great potential in the integration of the Web and social network search technologies. Such an integration can provide unified access to a significantly larger body of online information than what is currently available in the shallow Web. We presented evidence that the integration can also improve the quality of Web search results by ranking the results relative to the interests and biases of groups of individuals. In this section, we discuss research opportunities and challenges associated with realizing this vision.

Privacy: Participants in a social network must be willing to disclose which information they find interesting and relevant. This creates a tension between the privacy concerns of individuals and the effectiveness of the social network, which depends on the willingness of individuals to share information. In small social networks of mutually trusting participants (e.g., family members or close friends) the problem reduces to access control. However, in larger social networks (e.g., all researchers in computer networking), a solution that is acceptable to users would require mechanisms to control information flow and anonymity.

Membership and clustering of social networks: In general, an individual may be a participant in multiple social networks (e.g., networks related to professional interests, networks related to hobbies, and networks related to family and friends). This raises many questions. Are there automated mechanisms by which we can infer the social links between users? For instance, by observing email exchange between users, or by considering similarity in content browsed or stored between pairs of users. Similarly, can we automatically identify different clusters of communities associated with certain interests? In the absence of such techniques, users have to explicitly declare and manage their social network memberships. Finally, if a user participates in many social clusters, how should a search query be resolved with respect to the different clusters?

Content rating and ranking: The use of social network-based search techniques enables new approaches to ranking search results. There are many alternatives that could be explored: should we use global page rank, as in Google, or should we use a local page rank specific to the social network? Should content be ranked based on the number of users who have viewed or stored the content, or should the ranking be based on explicit user ratings of the content? Furthermore, how should the search results from the social network be displayed or ranked relative to the Google results?

System architecture: Should the system be centralized or distributed? A centralized architecture, similar to current Web search engines, may raise concerns about privacy, trust and market dominance. Also, a centralized approach may not scale with the bandwidth requirements of a central data store or the number of different social networks. A decentralized architecture, on the other hand, faces challenges of its own. Building even a conventional Web search engine in a decentralized fashion is an open research problem. Adding decentralized social network search requires scalable, index-based search algorithms, and appropriate mechanisms to ensure privacy.

5 RELATED WORK

Several projects have looked at replacing the functionality of the large centralized Web search engines with a decentralized system, built from contributing users' desktops [11]. Both Minerva [2] and YaCy [22] implement a peer-to-peer Web search engine without any points of centralization. Additionally, other projects [10, 19] have examined replacing the centralized PageRank computation of Google with a decentralized approach. All of these projects, though, are primarily focused on replacing the functionality of existing centralized search engines with a decentralized architecture.

A few systems have looked at query personalization, or taking a user's preferences and interests into account when ranking pages. Most notably, A9 [1] and Google Personalized Search [7] allow users to create profiles to which search results are tailored. There has also been much research into methods for accurately personalizing search queries [9, 21]. While these projects are concerned with personalization, our work is complementary and examines the ability to use social links to improve search results.

Lastly, a number of projects have looked at using social networks to aid a variety of applications. Notable distributed systems projects include SPROUT [14], which uses the trust of social links to increase the probability of successful DHT routing, and Maze [23], which allows users to create friends in the file sharing network.

6 CONCLUSION

In this paper, we examined the potential for using online social networks to enhance Internet search. We analyzed the differences between the Web and social networking systems in terms of the mechanisms they use to publish and locate useful information. We discussed the benefits of integrating the mechanisms for finding useful content in both the Web and social networks. Our initial results from a social networking experiment suggest that such an integration has the potential to improve the quality of Web search experience. Finally, we outlined research challenges in leveraging online social networks to build search systems for the future Internet.

REFERENCES

[1] A9 Search. `http://www.a9.com`.

[2] M. Bender, S. Michel, P. Triantafillou, G. Weikum, and C. Zimmer. MINERVA: Collaborative P2P search. In *Proceedings of the 31st International Conference on Very Large Data Bases (VLDB'05)*, August 2005.

[3] BlogSpot. `http://www.blogspot.com`.

[4] Flickr. `http://www.flickr.com`.

[5] FreePastry Project. `http://www.freepastry.org`.

[6] Friendster. `http://www.friendster.com`.

[7] Google Personalized Search. `http://www.google.com/psearch`.

[8] Google Video. `http://video.google.com`.

[9] G. Jeh and J. Widom. Scaling personalized web search. In *Proceedings of the 12th International World Wide Web Conference (WWW'03)*, May 2003.

[10] K.Sankaralingam, S. Sethumadhavan, and J.C.Browne. Distributed PageRank for P2P systems. In *Proceedings of the 12th IEEE International Symposium on High Performance Distributed Computing (HPDC-12)*, June 2003.

[11] J. Li, B. T. Loo, J. Hellerstein, F. Kaashoek, D. R. Karger, and R. Morris. On the feasibility of peer-to-peer web indexing and search. In *Proceedings of the 2nd International Workshop on Peer-to-Peer Systems (IPTPS'03)*, February 2003.

[12] LiveJournal. `http://www.livejournal.com`.

[13] Lucene Search Engine. `http://lucene.apache.org`.

[14] S. Marti, P. Ganesan, and H. Garcia-Molina. DHT routing using social links. In *Proceedings of the 3rd International Workshop on Peer-to-Peer Systems (IPTPS'04)*, February 2004.

[15] MySpace. `http://www.myspace.com`.

[16] MySpace is the number one website in the U.S. according to hitwise. HitWise Press Release, July, 11, 2006. `http://www.hitwise.com/press-center/hitwiseHS2004/social-networking-june-2006.php`.

[17] Orkut. `http://www.orkut.com`.

[18] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[19] J. X. Parreira, D. Donato, S. Michel, and G. Weikum. Efficient and decentralized PageRank approximation in a peer-to-peer web search network. In *Proceedings of the 32nd International Conference on Very Large Data Bases (VLDB'06)*, September 2006.

[20] B. Smyth, E. Balfe, O. Boydell, K. Bradley, P. Briggs, M. Coyle, and J. Freyne. A live-user evaluation of collaborative web search. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI'05)*, July 2005.

[21] J. Teevan, S. T. Dumais, and E. Horvitz. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th Annual International ACM SIGIR Conference (SIGIR'05)*, August 2005.

[22] YaCy Search Engine. `http://www.yacy.net`.

[23] M. Yang, H. Chen, B. Y. Zhao, Y. Dai, and Z. Zhang. Deployment of a large-scale peer-to-peer social network. In *Proceedings of the 1st Workshop on Real, Large Distributed Systems (WORLDS'04)*, December 2004.

[24] YouTube. `http://www.youtube.com`.