Title: Social Web Search

Name: Maryam Shoaran, Alex Thomo, Jens Weber

Affil./Addr.: Department of Computer Science

University of Victoria

PO Box 3055, STN CSC, Victoria, BC, Canada V8W 3P6

Phone: (250) 472-5786

Fax: (250) 472-5708

E-mail: {maryam, thomo, jens}@cs.uvic.ca

Social Web Search

Synonyms

social navigation, social analysis, social content search, recommender systems

Glossary

Social Media: online systems with high public participation and interaction rate.

User Metadata:. data created as the result of user interactions in an information space.

Socially Enhanced Search: quality improved search resulting from employing user metadata.

Personalization: adjustment of a system or process to fit user preferences.

Blogosphere: collection of interconnected Web logs.

Facebook Graph Search: information lookup in the Facebook graph-structured data.

Collaborative Filtering: discovery of new knowledge and patterns through filtering data produced by collaboration between different individuals.

Definition

Social search is an online search process that employs user-generated data and user-user relationships produced by social systems including bookmarking sites, web forums, social networks, and blogs to discover the best matching content to user queries in an information space. This is different from the methods used in traditional Web search engines in the sense that search techniques in the latter are mostly based on page-author-generated data such as page content, anchor text, and link connections. User-created data forms a rich source of metadata that expresses single-user or community preferences, ideas, and needs. User tags and queries can be considered as new descriptions of Web page content. Social search utilizes this new and fast expanding source of information to establish a fine-grained and more personalized or community-based online search.

A variety of information systems ranging from the World Wide Web to special purpose social systems, such as social networks, bookmarking sites, document or media sharing communities, and e-commerce benefit from the capabilities of social search. In the literature, social search also refers to the process of the analysis and discovery of new knowledge from social media.

Introduction

The objective of an online search system is to locate the relevant objects (e.g., Web pages) to a user-generated query from the Web or a community based collection. Over decades, Web search engines have improved their quality of search by inventing new techniques to retrieve query-relevant documents and rank them based on their quality. The characteristic of almost all of these techniques is that they are based on the data created by the Web page builders or document authors. Two types of ranking methods

are used in search engines. First, query-dependent or similarity measures that use document content, title, and anchor text to find similar documents. Second, query-independent or static measures that use page connectivity (link structure) as a quality measure to rank similar documents. The prominent static metrics are PageRank [18] and HITS [12].

Recently, with the ever-increasing activity and popularity of social media, a new type of information – user-created metadata – is available that can be used to enhance the quality of search.

User-generated content can be categorized as *explicit* or *implicit*. Explicit user data is created by visitors of web sites in the form of annotations and viewpoints in order to describe, organize, and share their favorite entities (URLs, movies, songs, books, articles, etc) online. Social systems capture explicit user annotations and viewpoints (feedback) in different forms. For example book, article, and movie review sites collect user reviews and ratings as text and star points. Social bookmarking sites store user tags and favorite URLs, and social networks capture user comments and their likes. User annotations and viewpoints constitute a precious information source that can be utilized to extract, for example, Web page descriptions (using tags and comments), page or media popularities (using bookmarks and ratings), and user preferences (using ratings and likes).

Monitoring user online behavior builds another valuable source of information. Implicit user data is automatically extracted from system logs containing user search queries, browsing history (clickthrough data), and amount of time spent by users on different pages. This data can help improve the quality of search in different ways. For instance, user queries can be considered as "URL tags" describing the content of pages. User browsing history is an indication of user interest and can be used to resolve the ambiguity that often exists in user queries. The amount of time spent by

users reading the content of websites might be an indication of the importance of sites and can be used to improve the ranking process, especially in community-based search environments.

Social Web Search and Analysis

Online social systems holding a rich public participation have been able to accumulate valuable and heterogeneous collections of user metadata. Social search and analysis is focused on taking advantage of such data sources by new techniques that either help to improve the functionality of already existing systems or devise novel analysis and knowledge discovery schemes. Social search and analysis is active in different areas as follows: (1) socially enhanced web search, (2) social navigation, (3) social analysis, (4) recommender systems, and (5) social content search.

Socially Enhanced Web Search

Social web search aims at improving the quality of web search by combining traditional search methods, e.g. query-document similarity and PageRank, with new techniques that employ social content. For instance, *SocialSimRank* (SSR) and *SocialPageRank* (SPR) [5] are two new methods that integrate social annotations available in social bookmarking sites, e.g. del.icio.us [1], into the page ranking process.

SocialSimRank (SSR) is a similarity ranking algorithm for queries and social annotations. The algorithm is based on the assumption that social annotations provide good summaries of web pages from various user perspectives. Based on this observation, similarities for every pair of annotations and similarities for every pair of pages are iteratively computed. The similarities are recursively defined as follows. The more similar the pages are, the more similar their corresponding annotations are. Conversely, the more similar annotations are, the more similar their associated pages are. These

similarities are integrated into each others computation. That is, in the equation that calculates the similarity between two annotations, one of the parameters is the similarity between two pages to which these annotations are assigned, and vice versa. After several iterations, this process typically converges, and the system is ready to answer queries. Each query term is considered to be a page annotation. The similarity of a query q to a web page p is computed as the sum of the similarities of each term in q to each annotation associated with p.

SocialPageRank (SPR) computes the page quality (popularity) with the intuition that the number of annotations assigned to a web page indicates the quality of the page in some sense. SPR uses an iterative algorithm to compute page popularities based on user and annotation popularities. Integrating SSR and SPR into a ranking function that also uses traditional document-similarity metric and PageRank improves the quality of web search [5].

Other enhanced search methods are also proposed that benefit from different aspects of user annotations. A hybrid search technique is presented in [21] that combines a link-based ranking method with a new metric that is based on user-generated data in social bookmarking sites (e.g. del.icio.us). The new metric utilizes SBRank (Social Bookmarking Rank) which is the number of user bookmarks on a page, the sentiment-based and temporal information extracted from user annotations, as well as general statistics derived from user interactions with web pages.

Community-based search systems improve the quality of search by incorporating user search behaviors within the community, e.g. user queries and result selections, into the ranking method. The underlying intuition is that among the users of similar mind, e.g. social network or enterprise intranet users, the context of queries is similar, the query repetition is high and also there rarely exist malicious behaviors that can negatively affect popularity metrics [8]. This type of social search is also called *Collabo-*

rative Web Search (CWS) [17] and I-SPY [20] is an implementation of it. Such systems record the queries and result selection of the community searchers and upon exposure to a new query the information of search sessions of a similar pattern is retrieved. The system re-ranks the result returned by the underlying search engines to reflect the implicit preferences of the community. Each item in the result list is also augmented by a set of past related queries that can be used to start new searches.

Social Navigation

The goal of social navigation is to enhance the quality of user browsing by providing various types of navigational assistance based on the visiting behavior of similar–minded users in the past. Social navigation systems benefit from different implicit and explicit user-generated data. They keep track of the browsing behavior of the users by collecting user queries and browsing paths (personal footprints). The time spent reading a page is also taken into account as an indication of user intention. Such systems also benefit from user annotations that can provide useful information about the importance of visited pages. When a user clicks on a source or on a (page) link in the search result list, the system provides a visual guide containing different navigational cues, for example, the source or page visit frequency (browsing popularity), the number of associated annotations (annotation popularity), and a list of queries leading to this source or page (search popularity).

Knowledge Sea II [6] is an example of a social browsing system that was developed to help students in a class to find the most useful sources for a particular course. This system organizes sources in a table with each cell associated with one source. The available navigational clues include the background color of cells indicating visit frequency, a sticky note for the presence of annotations, and a thermometer representing the number of positive annotations.

Another interesting system is the one presented in [8] that facilitates community-based access to the Communications of ACM (CACM) magazine. This system integrates social search and social navigation in both the interface level and its internal mechanisms. When a new search query is initiated, the search component of the system retrieves similar queries and their associated search results. Then, the results are scored based on their relevance to the new query, and finally the top-k results are placed ahead of the other results returned by the ACM search engine. Each result item is appended by complementary information presented as icons. Five icons with different levels of filling indicate, respectively, (1) the relevance of the result to the query (the percentage of times the result has been selected for the query by community users), (2) a list of other queries that have led to the selection of this result by community users, (3) the last time the result was encountered by the users (a view of the freshness), (4) the browsing popularity of the result (footprints), and (5) the user annotations. When a result is selected, the browsing component also augments the opened pages with social assistance icons.

Social Quality Analysis

The quality of user-generated content in online social systems varies from excellent to spam due to the participation of individuals with different intentions and levels of expertise. This is especially important in knowledge-based social systems such as question/answering portals, online forums, and networks of email exchangers. Social quality analysis aims at identifying knowledge experts and high quality user-created content in order to improve the quality of information-retrieval tasks (cf. [23; 7; 3; 22]). Various analysis methods are used ranging from link-based ranking algorithms, e.g. PageRank [18] and HITS [12], to text classification techniques and user-clickthrough information.

Since 2006, some interesting systems have been presented that automatically evaluate the quality of questions and answers in question/answering domains (cf. [11; 3]). The framework presented in [3] first identifies a collection of quality–indicating features of social media and associated interactions. Then, these features are used as input to a classifier (a stochastic gradient boosted tree), in order to extract high quality content. A wide range of information sources are used to extract features of the following categories. (1) Content-based: textual features of questions and answers, such as word n-grams, punctuation and typos, syntactic and semantic complexity measures, and grammaticality measures; (2) Connectivity-based: link-based metrics (authority scores and PageRank) in user-item and user-user relationship graphs, where an item is a query or an answer; (3) Usage-based: temporal statistics, number of clicks on items, and time spent on reading.

Recommender Systems

In online shopping, movie, and music web sites the goal is to improve the user experience by providing appropriate recommendations about new items that match user interests, ideas, and needs. These Web sites collect different types of user-produced data, ranging from explicit user ratings to implicit purchase history, browsing and search activities. Recommender systems [19], using sophisticated algorithms, combine data from independent contributors to discover new knowledge about relations between users and items.

There are two major approaches in recommender systems, *content filtering*, and *collaborative filtering* [15; 14]. Content filtering discovers matching users and items based on their individual characteristics. Items (products) are profiled by domain experts and user profiles are created by users' explicit answers to specific questions, e.g.

demographic questions. A problem with content based filtering is the difficulty of gathering relevant information.

Collaborative filtering, on the other hand, is based on user behavior in the past, for example user transactions and product ratings. By analyzing the relationships among users and among items, collaborative filtering predicts new relations between particular users and items. Suppose that in a movie rental site, user u has not watched and rated movie x yet, and the system would like to know whether it should recommend x to u. In the user-centered collaborative approach, first the similarity between u and all other users who have rated x is computed using some similarity measure, e.g. Euclidean distance or Pearson correlation coefficient. Then, the system predicts how u would rate x by computing a weighted average of the ratings for x by the most similar users to u. If the predicted rating is above a certain threshold, the system recommends x to u. In the item-centered approach the prediction is made based instead on the similarity between items.

Motivated by the Netflix prize contest significant improvements have been made in the quality of recommender systems. *Latent factor* models are another approach in collaborative filtering that maps the users and items to a common multidimensional space based on the past rating patterns. Latent factor models are based on *sparse matrix factorization*, and they are among the most popular and the best performing approaches [13; 15; 14].

Social Content Search

Despite the similarities between social media sites such as social networks, the blogosphere, and microblogging systems like Twitter they differ in the type of data predominantly posted and shared by users, as well as in the form of user interconnections they

offer. Based on these characteristics special-purpose search and information discovery efforts are applicable on the content of each site [2; 4; 16].

Facebook has recently launched *Graph Search* [2] as a new feature to benefit from its massive storage of data and relationships. Using this tool people can search for real world objects in Facebook's knowledge graph, which is comprised of objects such as people, places, and things and inter-object connections, for example Friendship and Likes. An important advantage of Facebook's search is that it has access to the collective knowledge of its vast community of users (more than one billion) to answer questions involving different layers of searching. Appealing examples are: "What to read that is liked by my friends in college", "Where to eat in Toronto that my friends living there like", "Where to go in Asia that my friends and friends-of-friends of my age found interesting", "What iPhone app to download that my friends use to track their jogging and cycling", etc. People's experience with Facebook Graph Search will highly depend on the level of their connectedness and participation in the system.

Blogging is another online social activity that has received an increasing popularity in recent years. The free context of blogs makes the Blogosphere (the collection of connected blogs) a rich source of heterogeneous information including personal experiences and opinions about a variety of subjects. Mining and analysis of blog data can capture public insight in different topics [9; 4]. For instance, BlogScope [4] is one of the systems designed to analyze the textual content of blogs and to provide information such as *when*, *where*, and *why* about interesting topics. When the user selects one of the daily hot keywords provided by the system or poses a query, all the relevant blog posts are retrieved and the result of various analysis on their content is presented. For example, the system can display the following information; (1) A popularity curve for a keyword as a function of time; (2) A list of the most closely related keywords in blog posts; (3) A distribution of the related posts on the map; (5) A synopsis set which is

the maximal set of keywords correlated with query that exhibits a bursty behavior in the associated popularity curve.

Future Directions

Whereas the usefulness of the annotations in small scaled information communities has been shown by several works, social annotations and bookmarks lack yet the sufficient size to significantly influence the performance of search engines in a large scale (cf. [10; 5]). Augmenting more Web sites with improved tagging systems that contain for example appealing and structured user interfaces and also provide incentives to stimulate tagging activities would significantly help to improve the situation.

Cross-references

Collaborations in Online Social Networks; Collective Intelligence; Microtext Processing; Recommender Systems, Models and Techniques; Social Bookmarking.

References

1. del.icio.us. https://delicious.com/.
2. Facebook graph search. https://www.facebook.com/about/graphsearch.
3. E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *WSDM*, pages 183–194, 2008.
4. N. Bansal and N. Koudas. Blogscope: spatio-temporal analysis of the blogosphere. In *WWW*, pages 1269–1270, 2007.
5. S. Bao, G.-R. Xue, X. Wu, Y. Yu, B. Fei, and Z. Su. Optimizing web search using social annotations. In *WWW*, pages 501–510, 2007.
6. P. Brusilovsky, G. Chavan, and R. Farzan. Social adaptive navigation support for open corpus electronic textbooks. In *AH*, pages 24–33, 2004.

7. C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *CIKM*, pages 528–531, 2003.
8. J. Freyne, R. Farzan, P. Brusilovsky, B. Smyth, and M. Coyle. Collecting community wisdom: integrating social search & social navigation. In *IUI*, pages 52–61, 2007.
9. D. Gruhl, R. V. Guha, R. Kumar, J. Novak, and A. Tomkins. The predictive power of online chatter. In *KDD*, pages 78–87, 2005.
10. P. Heymann, G. Koutrika, and H. Garcia-Molina. Can social bookmarking improve web search? In *WSDM*, pages 195–206, 2008.
11. J. Jeon, W. B. Croft, J. H. Lee, and S. Park. A framework to predict the quality of answers with non-textual features. In *SIGIR*, pages 228–235, 2006.
12. J. M. Kleinberg. Hubs, authorities, and communities. *ACM Comput. Surv.*, 31(4es):5, 1999.
13. Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *KDD*, pages 426–434, 2008.
14. Y. Koren and R. M. Bell. Advances in collaborative filtering. In *Recommender Systems Handbook*, pages 145–186. 2011.
15. Y. Koren, R. M. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *IEEE Computer*, 42(8):30–37, 2009.
16. M. Mathioudakis and N. Koudas. Twittermonitor: trend detection over the twitter stream. In *SIGMOD Conference*, pages 1155–1158, 2010.
17. M. R. Morris and J. Teevan. *Collaborative Web Search: Who, What, Where, When, and Why*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers, 2009.
18. L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999. Previous number = SIDL-WP-1999-0120.
19. P. Resnick and H. R. Varian. Recommender systems - introduction to the special section. *Commun. ACM*, 40(3):56–58, 1997.
20. B. Smyth, E. Balfe, J. Freyne, P. Briggs, M. Coyle, and O. Boydell. Exploiting query repetition and regularity in an adaptive community-based web search engine. *User Model. User-Adapt. Interact.*, 14(5):383–423, 2004.

21. Y. Yanbe, A. Jatowt, S. Nakamura, and K. Tanaka. Can social bookmarking enhance search in the web? In *JCDL*, pages 107–116, 2007.

22. L. Yang, S. Bao, Q. Lin, X. Wu, D. Han, Z. Su, and Y. Yu. Analyzing and predicting not-answered questions in community-based question answering services. In *AAAI*, 2011.

23. J. Zhang, M. S. Ackerman, and L. A. Adamic. Expertise networks in online communities: structure and algorithms. In *WWW*, pages 221–230, 2007.