The Social Future of Web Search: Modeling, Exploiting, and Searching Collaboratively Generated Content

Eugene Agichtein
Emory University
eugene@mathcs.emory.edu

Evgeniy Gabrilovich
Yahoo! Research
gabr@yahoo-inc.com

Hongyuan Zha
Georgia Institute of Technology
zha@cc.gatech.edu

Abstract

Social, or collaboratively generated content (CGC) is transforming how we seek and find information online: it is now a prominent part of the web information ecosystem, and a powerful platform for information seeking. The resulting archives of both the content and the context of the interactions contain valuable information that is often not available elsewhere, and can be helpful for the development of novel ranking algorithms, and natural language processing, text mining, and information retrieval techniques. We review machine learning techniques for modeling CGC, focusing on tasks such as learning to estimate content quality, relevance, and searcher intent and satisfaction with the retrieved results. We describe how this information can be incorporated into learning-based ranking methods for searching social media, and how CGC could be used to improve performance on key text mining and search tasks.

1 Introduction

Proliferation of the Internet and ubiquitous access to the Web enable millions of Web users to collaborate online on a variety of activities. Many of these activities result in the construction of large repositories of knowledge, either as their primary aim (e.g., Wikipedia) or as a by-product (e.g., Yahoo Answers). During the last few years, this user- or collaboratively-generated content (CGC) has started dominating the web: increasingly, users participate in content creation, rather than just consumption. Published and professional web content together is estimated to be generated at about 5 Gigabytes/day, whereas user-generated content is created at the rate of about 10 Gigabytes a day, and growing [RT08]. Popular user-generated content domains include blogs and web forums, folksonomies, social bookmarking sites, photo and video sharing communities, as well as social networking platforms such as Facebook and MySpace, which offer a combination of all of these with an emphasis on the relationships among the users of the community.

The unprecedented amounts of information in the collaboratively generated content sites, can enable new, knowledge-rich approaches to information access, which are significantly more powerful than the conventional word-based methods. Considerable progress has been made in this direction over the last few years. Examples include explicit manipulation of human-defined concepts and their use to augment the bag of words, using large-scale taxonomies of topics from Wikipedia or the Open Directory Project to construct additional class-based features, or using newly available word senses and examples of their usage for better disambiguation.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

At the same time, mining and organizing information in CGC is a challenging task. Unlike "traditional" web content, social media content is inherently dynamic: for hours or days after the initial posting of an item, users continue to post comments (e.g., on blogs and forums), vote for stories (e.g., on Digg), and contribute and rate answers to questions (e.g., on Yahoo! Answers). As responses and ratings arrive, the perceived popularity, quality, or even interpretation of an item may change significantly over time. Another important difference between collaboratively-generated and traditional content is the variance in the content quality. As Anderson [And06] describes, in traditional publishing—mediated by a publisher—the typical range of quality is substantially narrower than in niche, unmediated markets whereas social media content ranges from very high-quality items to low-quality, sometimes abusive content. This makes the tasks of filtering and ranking in CGC systems more complex. At the same time, social media presents inherent advantages over traditional collections of documents: the rich structure of CGC offers, in addition to document content and link structure, a wide variety of metadata about authorship, user feedback, as well as interactions with the content and other users.

Even the notion of "search" is not well understood in the social media context. For example, social media users often discover information by following recommendations from their social network (e.g., sites such as Facebook.com or Myspace.com), from other users on a site (e.g., sites such as Digg.com or Slashdot.org), by following some users explicitly (e.g., Twitter.com), or by contributing new content annotations or using existing tags contributed by other users (e.g., del.icio.us). These ways of finding and sharing information differ significantly from "traditional" keyword querying and relevance ranking models of web search. Thus, specialized techniques for modeling, analyzing, and searching in the social media environment have become an active area of research.

This paper surveys recent progress on modeling, exploiting, and searching collaboratively generated content. First, we review recent progress on modeling the process of social media creation, focusing on the user interactions, and the resulting quality of content. Then, we discuss how this content could be exploited to improve search-related tasks such as classification. Finally, we discuss recent work on incorporating this information to improve searching of social media and the web as a whole.

2 Modeling Collaboratively Generated Content

We first describe content creation in popular CGC sites, and some of the proposed models of content creation and evolution. Then, we review recent work on modeling content quality, both from the objective editorial perspective and that of user interest and engagement. These models are crucial for the work of exploiting and searching of social media content, described in later sections.

2.1 Content Creation

As representative examples, we consider Wikipedia, Youtube.com, and Yahoo! Answers web sites:

Wikipedia: Now a prototypical example of valuable collaboratively generated content, Wikipedia allows any user to create or edit virtually any page on the site, on any topic. The entire history of edits is preserved, allowing for reverting to an earlier version, and additional moderation is performed for controversial or popular topics. The central element of Wikipedia is an article, which serves as the unit of organization of the content, discussion comments, edits, and links. Numerous descriptive studies focused on Wikipedia content creation, including [CSC+06, Gil05, WH07].

Collaborative Question Answering (CQA): Community-driven question/answering portals enable users to answer questions posed by other users, providing an alternative to automatic web search: rather than submitting keyword queries to search engines, users express detailed information needs, and often obtain direct responses from other users. In some markets, notably in South Korea, this information seeking behavior has eclipsed the use of traditional web search. In the United States, Yahoo! Answers has attracted approximately 100 million users, and has a growing archive of more than 400 million answers to questions, according to 2008 estimates. The resulting content is increasingly incorporated into main search results by all of the major search engines. For concreteness, we focus on the Yahoo! Answers CQA portal, but the observations and procedures are quite

similar in other portals such as Naver and Baidu Knows. The central element of a CQA system are questions. Figure 1 shows a lifecycle of a question in CQA. It starts in an "open" state where it receives answers. Then at some point (decided by the asker, or by an automatic timeout in the system), the question is considered "closed", and can receive no further answers. At this stage, a "best" answer" is selected either by the asker or through a voting procedure by other users; once a best answer is chosen, the question is "resolved." The system is partially moderated by the community: any user may report another user's question or answer as violating the community guidelines (e.g., containing spam, adult-oriented content, copyrighted material, etc.). A user can also award a question a "star", marking it as an interesting question, and sometimes can vote for the best answer for a question, and can give to any answer a "thumbs up" or "thumbs down" rating, corresponding to a positive or negative vote, respectively. Recent studies describing the statistics of content generation in CQA include [ACD+08, AZBA08].

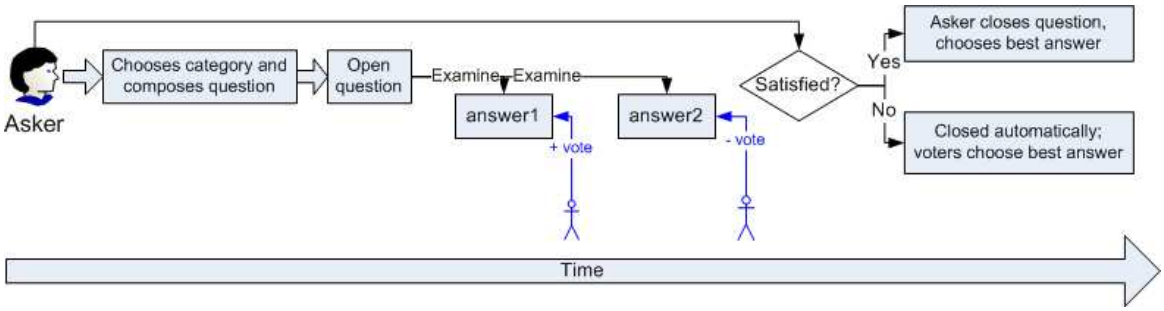

Figure 1: Question lifecycle in a typical CQA site

YouTube.com: As of 2009, Youtube.com is the largest user contributed video repository on the web, which allows comments, ratings, and annotations of the contributed videos by other users. Its central unit is a video clip, with many parallels to the CQA sites in that videos may be just a way to start a discussion amongst the users. Recent studies of Youtube.com content generation statistics include [CKR+07, SH08].

The proposed models for content creation and attention can be roughly divided into visibility-based and preferential-attachment models. In visibility-based models, such as the model proposed by Lerman [Ler07], users are more likely to view, rate, and contribute to recently posted items, with participation exponentially decaying over time. Not surprisingly, the study shows a strong connection between a contributor's social network size, and content popularity. However, in some important sites such as Wikipedia or CQA, the social network is fluid and ad-hoc: users post content and interact independently of their social network connections. Leskovec et al. examined micro-evolution of a number of social media sites, and empirically evaluated variants of a preferential attachment model [LBKT08]. More recently, a large-scale study of both YouTube and Digg introduced and evaluated three families of content dynamics models [SH08]. Interestingly, one of the proposed models (namely, *Growth Profile*) describes the user contributions with a uniform accrual curve, which is appropriately re-scaled to account for the differences between item interestingness—without making any assumptions often built into exponential decay or other parametric models. In summary, modeling CGC creation is a young and active research area, with more accurate and insightful models in development.

2.2 CGC Accuracy and Quality

Because of the high variance of content quality in social media, estimation of content quality is an essential module for performing advanced information retrieval tasks. For instance, a quality score can be used as input to ranking algorithms or used to more effectively exploit CGC for web search tasks. Recent work by Agichtein et al. [ACD+08] focused on the task of finding high quality content in social media. At the high level, their approach was to exploit features of social media that are intuitively correlated with quality, and then train a classier to select and weight the features for each specific type of item, task, and quality definition. Specifically, the authors modeled the intrinsic content quality, the interactions between content creators and other users, as

well as the content usage statistics. All of these feature types were then used as an input to a classifier that can be tuned for the quality definition for the particular media type. The three classes of features included:

- **Intrinsic (text) content quality**: word n-grams, punctuation, capitalization, and spacing density, as well as the fraction of misspellings and typos. Also included were features capturing syntactic and semantic complexity, text grammaticality (estimated using part-of-speech sequences), and statistical properties of term distributions compared to reference corpora.

- **Social network features**: A significant amount of quality information can be inferred from the relationships between users and items. In particular, link analysis-based features were used, such as the hub and authority scores [K97], as well as the PageRank score of each item [BP98].

- **Viewing statistics**: Both first order statistics such as counts of how many times the content was viewed, as well as normalization by topic/genre popularity and temporal characteristics such as item posting time and "age" were used.

The resulting models, trained over quality judgments contributed by experts, performed on par with human editors—that is, the resulting accuracy was comparable to the inter-editor agreement between experts. They investigated the contributions of the different sources of quality evidence, have shown that some of the sources are complementary (i.e., capture the same high-quality content using the different perspectives). The combination of several types of sources of information is likely to increase the classifier's robustness to spam, as an adversary would required to not only create content the deceives the classifier, but also simulate realistic user relationships or usage statistics.

2.3 CGC Popularity and Usefulness

Another important characteristic of collaboratively generated content is popularity. Recent work on predicting content popularity on sites such as Youtube.com [SH08] or Digg.com [Ler07] exploited both the temporal dynamics of the interactions, as well as the social network structure, demonstrating that it is possible to identify items that will eventually become popular, as early as 30 minutes or an hour after the initial posting.

In the context of CQA, a more appropriate metric is *user satisfaction*: whether an asker in CQA is *satisfied* with the answers contributed by the community [LBA08]. This is in contrast to the more traditional relevance-based assessments that are often done by judges different from the original information seeker, which may result in ratings that do not agree with the target user. Nowhere does the problem of subjective relevance arise more prominently than within CQA, where many of the questions are inherently subjective, complex, ill-formed, or often all of the above. Not surprisingly, user's previous interactions such as questions asked and ratings submitted are a significant factor for predicting satisfaction. We hypothesized that asker's satisfaction with contributed answers is largely determined by the asker expectations, prior knowledge and previous experience with using the CQA site.

The features used for prediction are similar to the editorial quality above, including the question text: words and 2-word phrases in the question, the wh-type (e.g., what or where), and the length of the subject (title) and detail (description) of the question; Question-Answer Relationship, including overlap between question and answer text, answer length, and the number of candidate answers; Community-ratings, including "thumbs up", "thumbs down", the asker's previous contribution history; the answerer reputation; and the statistical word distributions of the question and answer text.

A variety of classification algorithms were applied to learn to predict asker satisfaction with the obtained responses. Interestingly, the automatic system was significantly more accurate than human raters on this task. One possible explanation is that human raters were not able to take advantage of the context in which the question was asked, or the asker's expectations and prior experience with the site. The implications of these results are that in order by carefully modeling the CGC creation process, automatic techniques can be build to evaluate and filter this content with accuracy as high as, or higher, then human experts.

Having described how CGC is created, and the models of the resulting content quality and popularity, we now consider how this data could be used to improve search and information access tasks.

3 Exploiting CGC for Information Access

Since collaboratively generated content is contributed by humans, it naturally embodies large amounts of human knowledge about the world. Back in the early years of AI research, Buchanan & Feigenbaum [BF82] formulated the *knowledge as power hypothesis*, which postulated that "The power of an intelligent program to perform its task well depends primarily on the quantity and quality of knowledge it has about that task." When computer programs face tasks that require human-level intelligence, such as intelligent information retrieval, it is only natural to use repositories of human knowledge to endow the machines with the breadth of knowledge available to humans. In this section, we survey a number of techniques we recently developed for using collaboratively generated content to improve information access.

3.1 Explicit Semantic Analysis

Observe that repositories of collaboratively generated content usually contain textual information rather than highly structured knowledge such as, for example, CYC [LG90]. There are several ways to extract knowledge from CGC. One way to use such content is simply to view it as additional training data, which can be used in unsupervised or semi-supervised scenarios [AZ05, NMTM00, Joa99]. Alternatively, a transfer learning approach can be employed to learn from the labels (e.g., tags or categories) associated with the collaboratively generated content and then reuse the learned models across different but related learning tasks [BDH03, DN05]. Yet another, and perhaps most promising direction, is to use the knowledge encoded in CGC repositories in order to construct new features that enrich the language of text representation.

Feature generation techniques were found useful in a variety of machine learning tasks [MR02, Faw93, Mat91]. These techniques search for new features that describe the target concept better than the ones supplied with the training instances. The key observation is that many cases textual objects in CGC repositories are (explicitly or implicitly) associated with knowledge concepts. Examples of such concepts include nodes of the Open Directory (`dmoz.org`), titles of Wikipedia (`wikipedia.org`) articles, or tags in Flickr (`flickr.com`) or Del.icio.us (`del.icio.us`). To this end, we proposed to learn a *semantic interpreter* that identifies these concepts in input texts and quantifies the degree of affinity of each concept to the input text. Subsequently, new features corresponding to these concepts enrich (or even replace) the conventional bag-of-words representation of the input text. We call our method Explicit Semantic Analysis (ESA), as it uses knowledge concepts explicitly defined and manipulated by humans.

Let us illustrate the value of such knowledge with a couple of examples. Without using external knowledge (specifically, knowledge about financial markets), one can infer little information from a very brief news title "Bernanke takes charge". However, using the algorithm we developed for consulting Wikipedia, we find that the following concepts are highly relevant to the input: BEN BERNANKE, FEDERAL RESERVE, CHAIRMAN OF THE FEDERAL RESERVE, ALAN GREENSPAN (Bernanke's predecessor), MONETARISM (an economic theory of money supply and central banking), INFLATION and DEFLATION. As another example, consider the title "Apple patents a Tablet Mac". Without deep knowledge of hi-tech industry and gadgets, one finds it hard to predict the contents of the news item. Using Wikipedia, we identify the following related concepts: APPLE COMPUTER, MAC OS (the Macintosh operating system) LAPTOP (the general name for portable computers, of which Tablet Mac is a specific example), AQUA (the GUI of MAC OS X), IPOD (another prominent product by Apple), and APPLE NEWTON (the name of Apple's early personal digital assistant). In both cases, the concepts identified by ESA provide valuable relevant knowledge that can greatly help in interpreting and processing the original input text.

For ease of presentation, in the above examples we only showed a few Wikipedia concepts identified by ESA as the most relevant for the input. However, the essence is representing the meaning of text as a weighted combination of multiple knowledge concepts (e.g., *all* Wikipedia concepts). Then, depending on the nature of

the task at hand we either use these entire vectors of concepts, or use a few most relevant concepts to enrich the bag of words representation.

The ESA method imposes several simple requirements on a suitable knowledge repository:

1. It should be comprehensive enough to include concepts in a large variety of topics.
2. It should be constantly maintained so that new concepts can be promptly added as needed.
3. Since the ultimate goal is to interpret *natural* language, we would like the concepts to be *natural*, that is, concepts recognized and used by human beings.
4. Each concept should have associated text, which would serve as training example(s) for learning to recognize this concept in input texts.

Creating and maintaining such a set of natural concepts requires an enormous effort by many people, but fortunately this is exactly how CGC repositories are constructed and maintained. In an empirical evaluation, ESA-based semantic interpreters were implemented using two specific CGC repositories, namely, the Open Directory Project and Wikipedia. Using knowledge-rich features generated based on these repositories led to significant improvements in text categorization [GM07, GM09] and information retrieval [EGM08].

We believe the most important aspects of ESA are its ability to address synonymy and polysemy, which are arguably the two most important problems in natural language processing. Thus, two texts can discuss the same topic using different words, and the conventional bag of words approach will not be able to identify this commonality. On the other hand, the mere fact that the two texts contain the same word does not necessarily imply that they discuss the same topic, since that word could be used in the two texts in two different meanings. Our concept-based representation allows generalizations and refinements of word meaning, and partially address synonymy and polysemy.

The empirical evaluation showed that particularly large improvements in classifying short texts and in answering short queries. This behavior was expected, since external knowledge is required for interpreting short input texts correctly. The ESA technique also helped achieve state of the art results in computing semantic relatedness of natural language texts. In the extreme case—when computing relatedness of individual words—the bag of words approach is simply not applicable (except for the trivial case when the two words to be compared are identical). Consequently, to compute semantic relatedness of texts we completely replace their bag-of-words representation with the concept-based one. Notably, using a larger fraction of the available concepts led to a more fine-grained representation and consequently to a better estimation of semantic relatedness (as measured by higher correlation with gold-standard human judgments).

3.2 Domain-specific query augmentation using folksonomy tags

Folksonomy is a method for assigning user-defined labels to objects stored in public repositories of textual or multimedia content. Examples of popular folksonomies include FLICKR (a photo collection), DEL.ICIO.US (a bookmark sharing project), and YOUTUBE (a video sharing system). Typically, users can add tags to any object, whether they "own" it or not. Folksonomies facilitate interaction between Web users and promote knowledge sharing by integrating the user-defined tags in searching and browsing activities. In a sense, folksonomies comprise an alternative to restricted lexicons, as the numerous tags potentially allow users to achieve higher recall. When the original content creator might not have thought of all the applicable tags, users who subsequently encounter the object are likely to add tags they deem relevant.

In a recent study [BCG+09], co-tagging (i.e., tagging of the same object with different labels) was used to infer tag relatedness *in the context* of individual folksonomies. The key contribution was an alternative definition of context as a collection of tags assigned to or related to an object. Given a query, the system first identifies a set

of relevant tags, and then uses tags that co-occur with them to augment the query. This method performs domain-specific query disambiguation, and can actually learn that a query "menu" is likely to have food connotation on FLICKR but user interface connotation on DEL.ICIO.US.

This method was implemented for context-sensitive query augmentation using contextual advertising as an application domain. Note that a query submitted to FLICKR most likely conveys a different intent than the same query submitted to DEL.ICIO.US. That is, knowing at which site the query is submitted can help identify the user's search intent. Treating the content of the site as the context for queries, and matching ads accordingly, can potentially improve user experience. In the previous example, FLICKR ads for the query "menu" should ideally include offers from restaurants rather than services of UI experts, which would be more appropriate on DEL.ICIO.US. Having computed site-specific tag co-occurrence statistics, the relevant tags are used to expand the bag of words for the query, as well as classify those tags to create new taxonomy-based features. Thus, using CGC to represent queries in this rich feature space results in more relevant ad matches, so that the ads displayed on different folksonomy sites better reflect the intent of their users.

4 Searching Collaboratively Generated Content and the Web

To make CGC more accessible and serving a broad spectrum of users, it is important to provide an effective search interface. As discussed, CGC is different from the traditional content on the web in style, quality, and authorship. More importantly, the explicit support for social interactions between users, such as posting comments, rating content, and responding to questions and comments makes social media unique and requires new techniques for searching. In this section, we use searching collaborative question answering (CQA) archives as a concrete example to discuss the various issues involved in searching CGC, and the prominent problem of how to combine CGC and the generic Web in a unified search process.

4.1 Learning to Rank CGC

Question-Answering (henceforth QA) is a form of information retrieval where the users' information need is specified in the form of a natural language question, and the desired result is self-contained *answer* (not a list of documents). QA has been particularly amenable to social media, as it allows a potentially more effective alternative to web search by directly connecting users with the information needs to users willing to share the information. In CQA portals such as Yahoo! Answers and Naver, users can express specific information needs by posting questions, and get direct responses authored by other users.

At CQA sites both the questions and responses are stored for future use by allowing searchers to first attempt to locate an answer to their question, if the same or similar question has been answered in the past. As QA portals grow in size and popularity, searching for existing answers for a given question becomes increasingly crucial to avoid duplication, and save time and effort for the users. Finding relevant questions and answers of a new query in QA archives, however, is a difficult task that is distinct from web search and web question answering, exemplified by the availability of explicit user feedback and interactions, explicit authorship and attribution information, and organization of the content around topical categories and past question threads. Because of the increasing popularity of Yahoo! Answers, Naver and other CQA sites, several recent research efforts tried to address the CQA search problems. Jeon et al. [JCL05, JCLP06] presented retrieval methods based on machine translation models to find similar questions from Naver. An alternative approach is possible, based on the general framework of learning to rank, seamlessly integrating the interaction features into the question and answer retrieval [BLAZ08b, BLAZ08a]. The learning to rank approach used was based on gradient boosting [Fri01, ZZCS07]. In the following we focus on the specific issues in extracting features that incorporate user interactions.

The approach abstracts the social content in CQA sites as a set of QA pairs. Given a user query, the goal is to order the set of QA pairs according to their relevance to the query, and the ordering is done by learning a ranking function for triples $(query, question, answer)$. Users in CQA sites not only ask and answer questions, but also actively participate in regulating the system. A user can vote for answers of other users, mark interesting

questions and even report abusive behavior. Therefore, a CQA user has a threefold role: asker, answerer and evaluator. And there are respectively three types of user interaction activities: asking, answering and evaluating. Following the general practice in information retrieval, we can represent each query-question-answer triple $(query, question, answer)$ as a combination of textual features (i.e., textual similarity between query, question and answers), statistical features (i.e., independent features for query, question and answers) and social features (i.e., user interaction activities and community-based elements). In CQA sites, there is an additional important type of user feedback — user evaluation in the form of votes (represented as the "thumbs up" and "thumbs down" metaphors). We can use this information to infer preference relevance judgments for the set of answers.

We now elaborate on the social features: For each user in the user community of a CQA site, there are several features to describe his or her activities, such as the number of questions he or she asked, the number of answers he or she posted, the number of best answers he or she posted etc. These features to certain extent can approximate the user's expertise in the CQA site. And user's expertise within certain topics can in turn indicate the quality of his or her answers to the questions about certain topics. For example, in a query-question-answer triple, if answerer tend to post useful answers or even best answers in the past, he or she is more likely to give answers of high quality this time. Similarly reputation of askers and evaluators can also indicate quality of answers. Therefore, in each query-question-answer triple, we also extract features indicating user's activities in the CQA site such as "number of questions the asker asked in the community", "number of best answers the answerer posted in community", etc.

Empirical evaluations using Yahoo! Answers showed that textual features and community features are crucial, and that user feedback, while noisy, provides sufficient relevance information to improve the learning of the ranking functions. Interestingly, textual features were found less important for Precision at 1. The learning to rank method coupled with user interaction features achieves a significant improvement on the performance of QA retrieval over the Yahoo! Answers' default ranking and the supported optional votes-based ranking.

We emphasized that CQA sites heavily rely on active user participation, such as voting or rating of responses to a question. This user feedback is invaluable for ranking, filtering, and retrieving high quality content as we discussed above. Unfortunately, as collaborative question answering, and CGC in general, move into the mainstream and gain in popularity, the quality of the user feedback degrades. Some of this is due to noise, but, increasingly, a small fraction of malicious users are trying to "game the system" by selectively promoting or demoting content for profit, or fun. Hence, an effective ranking of CGC content must be robust to noise in the user interactions, and in particular to vote spam. Bian et al. [BLAZ08a] extend the learning to rank idea and consider general vote spam attack models. In particular, a new method of training a ranker was introduced, in order to increase its robustness to common vote spam attacks. The results of a large-scale experimental evaluation show that such a ranker is significantly more robust to vote spam, compared to a state-of-the-art baseline, as well as the ranker not explicitly trained to handle malicious interactions.

4.2 Unified ranking of CGC and Web documents

CGC and the Web are often complementary. It would benefit the users if a single search interface can be developed which provide a unified ranking of the contents of both. To this end, we consider a learning to rank framework that merges general Web search results with CGC results. One key observation is that the two ranking problems have very different data distributions and they also use different feature sets. For example, user interaction activities and community-based elements such as user votes are not available for Web search results. Therefore we can't rely on commonality between the learning tasks other than 1) they both rank documents by assigning scores to documents, and 2) that their *relevance labels* are comparable.

We advocate the approach to *simultaneously* learn two ranking functions, $h_W(\cdot)$ for Web documents and $h_A(\cdot)$ for CGC. We add preference constraints between documents across different rankings. These constraints usually involve only documents related to the same queries. With our assumption on consistent labels, they can be induced by these labels. For example, a *Good* document in Web ranking list one should be ranked higher than a *Fair* document in CGC ranking list. Suppose x and y are feature vectors of two documents related to the

same query but are in the ranking lists of Web search and CQA, respectively. When training h_W and h_A, we add constraint $h_W(x) \geq h_A(y)$, if we want to rank x higher than y in the merged lists. This requires that the learning algorithms for h_W and h_A be able to take this kind of constraints into account, or soft versions of such constraints using a gradient boosting-based ranking algorithm [Fri01, ZZCS07]. In addition to merging result lists, significant work is required to group and display the generic and social media search results appropriately and intuitively, which remains an active research area as new user models for search in social media emerge.

5 Conclusions

We believe that collaboratively generated content (CGC) sites, both as the platforms for information seeking and sharing, and as the source of human knowledge, have the potential to revolutionize information access. Accomplishing this requires new methods for modeling and processing this content, as well as techniques for incorporating information from CGC into search. In this article we presented a brief overview of recent progress on these exciting research directions.

ACKNOWLEDGMENTS: The first author acknowledges generous support from the Yahoo! Faculty grant, the Cisco "Improving the Human Network" program, and the Emory College Seed Fund.

References

[ACD+08] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media with an application to community-based question answering. In *Proceedings of WSDM*, 2008.

[And06] Chris Anderson. *The Long Tail: Why the Future of Business Is Selling Less of More.* Hyperion, 2006.

[AZ05] Rie Kubota Ando and Tong Zhang. Framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, pages 1817–1853, 2005.

[AZBA08] Lada A. Adamic, Jun Zhang, Eytan Bakshy, and Mark S. Ackerman. Knowledge sharing and yahoo answers: everyone knows something. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 665–674, 2008.

[BP98] Sergey Brin and Larry Page. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the 7th International World Wide Web Conference*, pages 107–117, 1998.

[BCG+09] Andrei Broder, Peter Ciccolo, Evgeniy Gabrilovich, Vanja Josifovski, Donald Metzler, Lance Riedel, and Jeffrey Yuan. Online expansion of rare queries for sponsored search. In *Proceedings of the 18th International World Wide Web Conference*, 2009.

[BDH03] Paul N. Bennett, Susan T. Dumais, and Eric Horvitz. Inductive transfer for text classification using generalized reliability indicators. In *Proceedings of the ICML-2003 Workshop on The Continuum from Labeled to Unlabeled Data*, 2003.

[BF82] B. G. Buchanan and Edward Feigenbaum. Forward. In R. Davis and Douglas Lenat, editors, *Knowledge-Based Systems in Artificial Intelligence.* McGraw-Hill, 1982.

[BLAZ08a] J. Bian, Y. Liu, E. Agichtein, and H. Zha. A few bad votes too many? towards robust ranking in social media. In *The 4th International Workshop on Adversarial Information Retrieval on the Web*, 2008.

[BLAZ08b] J. Bian, Y. Liu, E. Agichtein, and H. Zha. Finding the right facts in the crowd: Factoid question answering over social media. In *Proc. of 17th International World Wide Web Conference (WWW2008)*, 2008.

[CKR+07] Meeyoung Cha, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. In *IMC '07: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 1–14, 2007.

[CSC+06] A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: the case of wikipedia, 2006.

[DN05] Chuong Do and Andrew Ng. Transfer learning for text classification. In *Proceedings of Neural Information Processing Systems (NIPS)*, 2005.

[EGM08] Ofer Egozi, Evgeniy Gabrilovich, and Shaul Markovitch. Concept-based feature generation and selection for information retrieval. In *AAAI'08*, July 2008.

[Faw93] Tom Fawcett. *Feature Discovery for Problem Solving Systems*. PhD thesis, UMass, May 1993.

[Fri01] J. Friedman. Greedy function approximation: a gradient boosting machine. In *Ann. Statist.*, 2001.

[Gil05] Jim Giles. Internet encyclopaedias go head to head. *Nature*, 438:900–901, 2005.

[GM07] Evgeniy Gabrilovich and Shaul Markovitch. Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization. *Journal of Machine Learning Research*, 8:2297–2345, October 2007.

[GM09] Evgeniy Gabrilovich and Shaul Markovitch. Wikipedia-based semantic interpretation for natural language processing. *Journal of Artificial Intelligence Research*, pages 443–498, 2009.

[JCL05] J. Jeon, W.B. Croft, and J.H. Lee. Finding similar questions in large question and answer archives. In *Proceedings of CIKM*, 2005.

[JCLP06] J. Jeon, W.B. Croft, J.H. Lee, and S. Park. A framework to predict the quality of answers with non-textual features. In *Proceedings of SIGIR*, 2006.

[Joa99] Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the 13th International Conference on Machine Learning*, 1999.

[K97] Jon Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5):604–632, 1997.

[LBA08] Yandong Liu, Jiang Bian, and Eugene Agichtein. Predicting information seeker satisfaction in community question answering. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2008)*, 2008.

[LBKT08] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, 2008.

[Ler07] Kristina Lerman. Social information processing in social news aggregation. *IEEE Internet Computing: special issue on Social Search*, 2007.

[LG90] Douglas Lenat and R. Guha. *Building Large Knowledge Based Systems*. Addison Wesley, 1990.

[Mat91] Christopher J. Matheus. The need for constructive induction. In L.A. Birnbaum and G.C. Collins, editors, *Proceedings of the Eighth International Workshop on Machine Learning*, pages 173–177, 1991.

[MR02] Shaul Markovitch and Danny Rosenstein. Feature generation using general constructor functions. *Machine Learning*, 49(1):59–98, 2002.

[NMTM00] Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2-3):103–134, 2000.

[RT08] Raghu Ramakrishnan and Andrew Tomkins. Toward a PeopleWeb. *Computer*, 40(8):63–72, August 2007.

[SH08] Gábor Szabó and Bernardo A. Huberman. Predicting the popularity of online content. *CoRR*, 2008.

[WH07] Dennis M. Wilkinson and Bernardo A. Huberman. Cooperation and quality in wikipedia. In *WikiSym '07: Proceedings of the 2007 international symposium on Wikis*, pages 157–164, 2007.

[ZZCS07] Z. Zheng, H. Zha, K. Chen, and G. Sun. A regression framework for learning ranking functions using relative relevance judgments. In *Proc. of SIGIR*, 2007.