Research Statement

Hongning Wang
Department of Computer Science
University of Illinois at Urbana-Champaign

1 Research Overview

Recent years have witnessed the boom of big data, where large amounts of data are analyzed for knowledge discovery. The mined knowledge can be leveraged to optimize operations and decision making in many application domains, e.g., improving health care, enhancing homeland security, and accelerating scientific discovery. It is worth noting that a large portion of data is produced and consumed by humans, and it will continue to fuel exponential growth. For example, every day there are 5 million transactions processed on eBay [Answers, 2013], 340 million tweets posted on Twitter [Blog, 2012], and 5.13 billion searches performed on Google [Brain, 2013]. Decision makers - from ordinary users who make daily purchase decisions to senior officers in international organizations who make strategic plans - would all benefit from having access to the knowledge discovered from such human-generated data.

Comparing to the non-human-generated data, e.g., sensor data, internet traffic data and bio-sequence data, which is usually associated with predefined schemes and structures, human-generated data is mostly unstructured and thus not directly accessible by machines, posing significant challenges for data mining. Yet, human-generated data contains rich semantic information and may reflect the latent intentions of people, thus they are remarkably valuable for knowledge discovery. In order to effectively extract actionable knowledge from such human-generated data and facilitate end users exploiting the mined knowledge, we need to put human into the loop of mining big data via proper user modeling. Figure 1 illustrates my vision of human-centric data mining and knowledge discovery, where humans play a dual role as both data producers and knowledge consumers. First, as the producer of data, humans generate two
kinds of unique data, both are extremely valuable from the perspective of knowledge discovery: 1) As a consequence of human communications in natural language, people routinely generate enormous amounts of unstructured natural language text data; 2) As users of knowledge service systems, people interactively create constantly growing behavior data, e.g., search logs of a Web search engine. Properly modeling the underlying data generation process is necessary for discovering knowledge from such data. Second, as the consumer of the mined knowledge, different users have distinct preferences of information for decision making, thus accurately understanding their respective information needs and decision preferences is crucial for providing effective decision support. Blindly mining patterns or statistics from data will inevitably hide the big picture from us. As a result, the depth of user understanding and modeling directly determines the upper bound of optimality that the automated mining algorithms can achieve in mining the human-generated data and providing effective decision support. Raising this upper bound by developing effective computational models to analyze human-generated data is the general goal of my research.

Thus my research generally lies in the intersection between data mining and user modeling. I am especially interested in user understanding, knowledge discovery and decision optimization, via designing novel computational models. My works have result in over 20 publications in top venues in data mining and information retrieval areas, including SIGIR, KDD, WWW, WSDM and CIKM. I have explored a wide variety of human-generated data: including both the text data of opinionated reviews [Wang et al., 2010, Wang et al., 2011a], forum discussions [Wang et al., 2011b, Lu et al., 2012] and scientific literature [Wang et al., 2008, Wang et al., 2009], and user behavior data recorded in search engine logs [Wang et al., 2013c, Wang et al., 2013b]. The mined knowledge reveals users’ deep intention and behavior patterns, which provide unique signals for knowledge service system optimization and personalization [Wang et al., 2012, Wang et al., 2013a]. The models and algorithms I have developed enabled multiple real-life applications including, e.g., public opinion tracking and prediction [Wang et al., 2010, Wang et al., 2011a], retrieval system optimization [Wang et al., 2013c, Wang et al., 2013b], personalized information filtering [Wang et al., 2012, Wang et al., 2013a], and the public biological literature mining system ONBIRES [Huang et al., 2008].

2 Current Achievements

My Ph.D. research aims at developing general and effective computational models to: 1) discover knowledge via broad exploration of human-generated data, which includes opinionated review text data, online forum discussions and scientific literature, and user search behaviors recorded in search engine logs; 2) optimize the automated knowledge service systems’ output via the learned user models when interacting with the users.

2.1 Mining Human-Generated Data for Knowledge Discovery

In this line of work, I have developed novel methods to mine three different types of human-generated data as further discussed below.

Mining Opinionated Text Data: A vast amount of opinionated text data is available online, especially due to the dramatic growth of social media. Tapping into such data to obtain
knowledge about opinions of an entity such as a product, a person, a service, or a policy is an exciting emerging trend. To enable deep understanding of opinionated text data, I proposed and studied a new opinionated text mining problem called Latent Aspect Rating Analysis (LARA) in [Wang et al., 2010, Wang et al., 2011a]. Clearly distinct from all previous works in opinion analysis that mostly focus on integrated entity-level opinions, LARA for the first time reveals individual users’ latent sentiment preferences at the level of topical aspects in an unsupervised manner. Discovering such detailed user preferences (which are often hard to obtain by a human from simply reading many reviews) enables applications beyond traditional sentiment analysis studies. First, such analysis facilitates in-depth understanding of user intents. For example, by mining the product reviews, LARA recognizes which aspect influences a particular user’s purchase decision the most. Second, by identifying each user’s latent aspect preference in a particular domain (e.g., hotel booking), personalized result ranking and recommendation can be achieved. Third, discovering the general population’s sentiment preferences over different aspects of a particular product or service provides a more effective way for businesses to manage their customer relationship and conduct market research. My work in this direction has been well recognized by the research community (with 86 citations of my KDD’10 paper according to Google Scholar), and several commercial review websites (e.g., Expeida.com) have shown their interest in commercializing this technique.

I also studied the problem of organizing scattered online opinions by exploiting structured ontology [Lu et al., 2010]. Important aspects are automatically selected from the manually crafted ontology to organize opinions, while coverage and readability are optimized for generating the structured summaries. Going beyond analyzing each single user’s opinions, I further explored the problem of discovering opinion networks of users in social media [Lu et al., 2012]. Signals from both textual content (e.g., who says what) and social interactions (e.g., who talks to whom) are explored to identify groups of users who are strongly for or against each other on some topics.

**Mining User Behavior Data:** Accurately interpreting users’ preferences and behavior is necessary for optimizing a system’s service to users. To understand users’ longitudinal information seeking behaviors, I explored the user interaction patterns recorded in search engine logs (e.g., their issued queries and clicked documents). A novel learning-based method is developed to associate users’ separated search behaviors over time to reflect their underlying search intents [Wang et al., 2013b]. Capitalizing on the unique patterns of users’ search behaviors identified in this work, the proposed method is able to achieve performance gain while reducing the reliance on costly human annotations. This method has been deployed in the commercial search engine Bing’s internal tool chain; both researchers and engineers in Bing can now use this method for analyzing search logs.

Users’ result clicks have long been recognized as important signals for optimizing search result ranking; however, they are also known to be noisy and heavily biased [Joachims et al., 2005]. I developed a principled click modeling framework to estimate relevance quality of documents from noisy clicks [Wang et al., 2013c]. By explicitly modeling document content and dependency among users’ clicks, this method can estimate the relevance of documents with fewer or even no clicks. In contrast, no existing click models can make estimations in such a situation, because they only explore simple click statistics and require a significant volume of click history. This method has been used in the Yahoo! Shopping portal, where the system needs to promptly estimate the relevance of candidate items that may only associate with a handful of clicks.

**Mining Scientific Literature:** Extraction of entity relations, e.g., protein-protein interac-
tion in biological literature, is crucial to enable biomedical researchers to efficiently access frontier knowledge and thus accelerate scientific discovery. I have addressed three key issues related to the problem of automatic extraction of protein-protein interactions from biological literatures: namely filtering irrelevant documents by selecting and integrating rich types of features [Wang et al., 2008], extracting protein-protein interactions by pattern learning and validation [Huang et al., 2008], and identifying interaction detection methods used in the literature [Wang et al., 2009]. All these methods have been incorporated into the public biological literature mining system ONBIRES [Huang et al., 2006].

2.2 Optimizing Knowledge Service Systems via the Learned User Models

Discovering knowledge is never the only purpose of mining big data; supporting decision making for the end users with the mined knowledge is our ultimate goal. Based on the computational user models I have developed, I further studied the problem of optimizing knowledge service systems’ output according to different users’ information needs and decision preferences.

**Vertical Search Optimization:** Vertical search is characterized by its increasing importance of aspects beyond pure topical relatedness, e.g., freshness in news search, and price and credibility in product search. In the application of news search, I proposed a unified framework for modeling document freshness and topical relevance, as well as their relative importance, by query log and content analysis [Wang et al., 2012]. The method has been evaluated in the Yahoo! News search portal and achieved significant improvement against the originally deployed ranking algorithm. This framework is generally applicable to many other vertical search scenarios, and it opens up a new direction for optimizing search utility beyond the traditional notion of topical relevance.

**Personalized Web Search:** The common practice in modern Web search engines that train and apply a single ranking model across all users can hardly satisfy the diverse result ranking requirements from different users. To address this deficiency, I proposed to adapt the global ranking model for each individual user [Wang et al., 2013a]. The unique advantage of the method is its adaptation efficiency: the adaptation is achieved via a series of simple linear transformations on the global model according to each user’s click feedback, which enables online adaptation on a per-user basis. In comparison, most existing personalization methods require considerable amount of user history data, which prohibits their use in real-time scenarios.

**Opinion Analytic System:** Based on the techniques I have developed in my work of LARA [Wang et al., 2010, Wang et al., 2011a], I built a system named ReviewMiner\(^1\) for automated opinion mining and business intelligence. ReviewMiner provides aspect-based opinion analysis, and it automatically adapts to different users’ aspect preferences from their usage history in the system to perform personalized recommendation and ranking. Currently, no commercial review sites, e.g., Amazon (www.amazon.com) or Newegg (www.newegg.com), can provide such in-depth analysis of user opinions and preferences when assisting users to make purchase decisions. In addition, ReviewMiner provides functionalities for business analytic researchers to keep track of customer feedback and to understand customer opinions of products and services. For example, ReviewMiner can recognize the inquired item’s mostly commented aspects in the customer reviews, identify the corresponding relative emphasis the users have expressed over those aspects and track the temporal dynamics of user opinions and emphasis over those aspects. Such analysis can hardly be achieved in any other existing opinion mining or business analytic systems.

\(^1\)http://timan100.cs.uiuc.edu:8080/test-app/
3 Future Directions

As illustrated in Figure 1, my current research has mostly focused on mining human-generated data for knowledge discovery and decision support. My long-term research goal is to advance the level of user understanding and develop mining techniques to extract actionable knowledge from all types of data at scale for effective decision support. The ideal outcome of my research is an intelligent knowledge service system which directly *talks to the users and learns from the users*. Such a system is beyond the traditional retrieval systems (e.g., the commercial Web search engines of Bing and Google), but it is a fusion of systems for information extraction, integration, retrieval and operation optimization. The system would record every action the user has taken when interacting with the system (e.g., acquiring knowledge, browsing results, and posting contents), correctly responds to the user’s inquiry by analyzing his or her intention at that particular time, and actively updates its service strategy when the user’s interest changes. With such an intelligent system, users will have no need to be trained for any specific interaction scheme or to exhaust the combination of query keywords to describe their information need to the system (as they have to do in the current generation of information service systems). Instead, they could easily obtain the needed knowledge for all kinds of tasks through the support enabled by the deployed techniques of data mining and user intent modeling.

To achieve such a long-term goal, there are three major challenges to be solved: namely, knowledge discovery and representation, user understanding, and intelligent knowledge service techniques. My current Ph.D. research has prepared a solid basis for me to pursue such a long-term goal, and in my future research I plan to address all the three major challenges.

**Knowledge Discovery and Representation:** The types of data that convey invaluable knowledge are far beyond those I have explored in my Ph.D. research. Exploring all these types of data provides us a broad and comprehensive view in decision making and operation optimization. For example, mining personal health data helps doctors find more effective and affordable treatments for each patient; analyzing vehicle GPS sensor data can help alleviating public traffic congestion; and social media data provides a unique mine for sociologists to study social influence and dynamics. In my future research, with the collaboration with researchers in multiple disciplines, e.g., biomedical science, cyber-physical science and social science, I plan to further expand the mining techniques I have developed to cover more types of important data for knowledge discovery.

However, analyzing data in an ad-hoc manner is far from ideal, which will inevitably isolate the mined knowledge and hamper its usage across different applications. As all human-generated data is governed by the underlying intent of human behaviors, there are associations among different types of such data and all mining algorithms should be aware of these associations when modeling the data. As in my work of LARA, the latent user intents are abstracted as his or her emphasis over aspects of items in a specific domain. By knowing a user’s preference in one domain, we could infer his or her preferences in similar domains, e.g., preferring a lower price in hotel booking might indicate the emphasis of price in car rental and flight tickets as well. Along the line of exploring richer types of human-generated data, I plan to develop a general knowledge representation scheme, e.g., an ontology structure, to organize the mined knowledge about human behaviors, so that the learned knowledge can be used in multiple application scenarios transparently.

**User Understanding:** Problems related to human activities have their unique properties:
the human behaviors are largely determined by their own goals and preferences, rather than the instructions of system designers. As a result, effectively mining actionable knowledge from human activities requires a fusion of solutions from both computational methodologies and behavioural and psychology studies (e.g., human-computer interaction). In my current research, I have mainly concentrated on developing data-driven computational methods for user understanding. It would be fascinating to seek evidence and support for those computational methods from the behavioural and psychology studies, and incorporate the established theories from those fields to guide model building and knowledge mining. Such a study will never be just a one-sided effort; the developed method and mined knowledge about users will also benefit the studies in psychology, cognitive science, design, and more.

In addition, nowadays each individual’s behavior is no longer “individual,” but a result of social interactions. Therefore, in order to understand users in a more precise and comprehensive view, we should no longer treat them independently, but put our analysis into the context of social interactions. Based on my previous study in social opinion analysis, I plan to further explore user behaviors and interactions within their social networks in a broader view to advance the level of user understanding.

**Intelligent Knowledge Service Techniques:** Currently, most knowledge service systems respond to users’ inquires in a passive way, e.g., predefined service strategies or offline trained prediction models are used to generate the output. Such a strategy is only sub-optimal given that it cannot adapt to individual user’s information need or the shift of public interest quickly. In my blueprint of the intelligent knowledge service system, the system should not only quietly assist the users, but also actively learn from the users (e.g., probe users for additional feedback). That is, let the system and users collaborate for knowledge acquisition and decision making. Game theory provides a nice theoretic foundation to perform such study, but the research on the border of computer science and game theory is still in its infancy. The real problems in mining big data for knowledge discovery are complex, dynamic, and involve a lot of uncertainty; thus, it will not be a straightforward practice of applying existing solutions in game theory to those real-life problems. Most of the research that has been conducted in the past focused on restrict assumptions about the game and players; little is known about more general settings. In the future, I would like to explore the application of game-theoretic models for building the system that actively talks to the users and learns from the users in an intelligent way.

As a passionate computer science researcher, I am enthusiastic to contribute to this new and exciting research discipline of mining big data for human-centric knowledge discovery and decision optimization. I am eager to conduct independent innovative research and interdisciplinary collaboration with researchers in the areas of systems, networks, human computer interaction, psychology, business and more, and seek funding opportunities from multiple funding agencies and industries. I have the passion and confidence to carry out my proposed frontier research, contribute to the field and benefit the real world.

**References**


