Special Issue: Web Data Mining

Guest Editorial

Richard Khoury
Department of Software Engineering, Lakehead University, Thunder Bay, Canada
Email: rkhoury@lakeheadu.ca

The Internet is a massive and continuously-growing source of data and information on subjects ranging from breaking news to personal anecdotes to objective documentation. It has proven to be a hugely beneficial resource for researchers, giving them a source of free, up-to-date, real-world data that can be used in a varied range of projects and applications. Hundreds of new algorithms and systems are being proposed each year to filter out desired information from this seemingly endless amount of data, to clean and organize it, to infer knowledge from it and to act on this knowledge.

The importance of the web in scientific research today can be informally gauged by counting the number of published papers that use the name of a major website in their titles, abstracts, or keyword lists. To illustrate, we gathered these statistics using the IEEE Xplore search system for eight well-known websites for each of the past 10 years. The results of this survey, presented in Figure 1, indicate that research interest for web data is increasing steadily. The individual websites' naturally know increases and decreases in popularity; for example, we can see Google overtake Yahoo! around 2007. In 2011, the three most cited websites in our sample of the scientific literature were Google, Facebook and Twitter. This ranking is similar, but a bit off, compared to the real-world popularity of these websites as measured by the website ranking site Alexa. Alexa's ranking does put Google and Facebook in first and second place respectively, but rank Twitter ninth, below YouTube, Yahoo!, Baidu and Wikipedia. There is nonetheless a good similarity between the rankings of Figure 1 and those of Alexa, which is not unexpected: to be useful for scientific research a site needs to contain a lot of data, which means that it must be visited and contributed to by a lot of users, which in turn means a high Alexa rating.

This special issue is thus dedicated to the topic of Web Data Mining. We attempted to compile papers that touch upon both a variety of websites and a variety of data mining challenges. Clearly it would be impossible to create a representative sample of all data mining tasks and all websites in use in the literature. However, after thorough peer-reviewing and careful deliberation, we have selected the following papers as good examples of a range of web data mining challenges being addressed

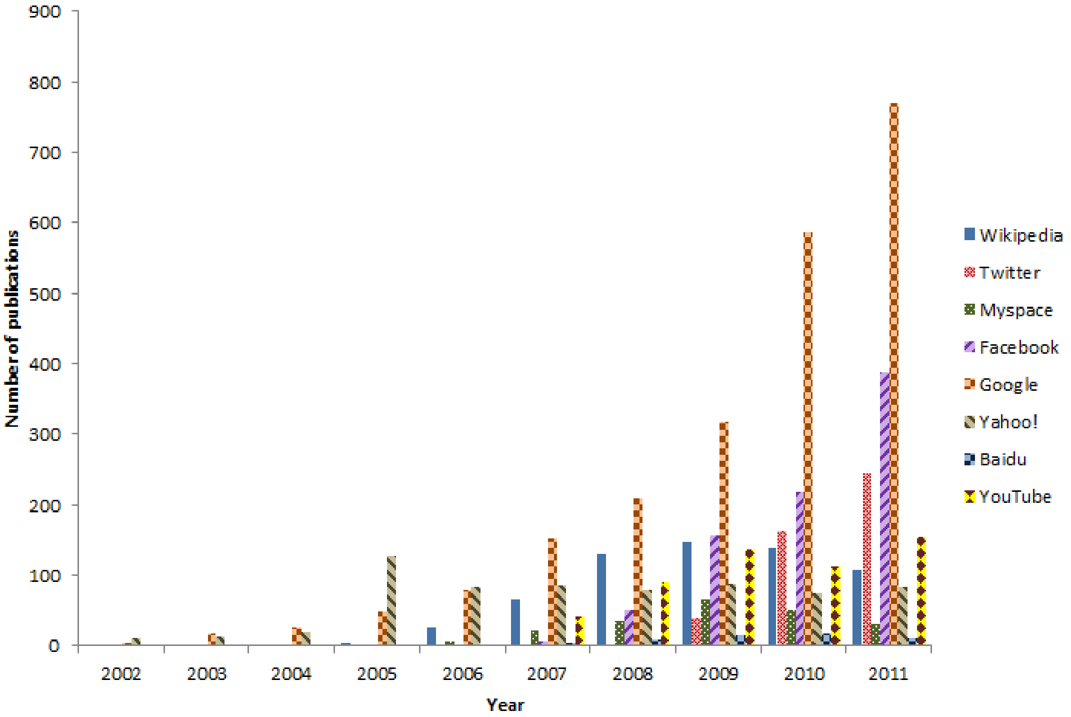

Figure 1. Number of publications per year that uses the name of a major website.

doi:10.4304/jetwi.4.3.205-206

today.

In our first paper, "Query Classification Using Wikipedia's Category Graph", the authors *Milad AlemZadeh*, *Richard Khoury* and *Fakhri Karray*, perform data mining on Wikipedia, one of the more popular websites both in the literature and to the public. The task they focused on is query classification, or the challenge of determining the topic intended by a query given only the words of that query. This is a challenge with broad applicability, from web search engines to question-answering systems. Their work demonstrates how a system exploiting web data can perform this task with virtually no domain restrictions, making it very appealing for applications that need to interact with human beings in any setting whatsoever.

Next, we move from Wikipedia to Twitter, a website whose popularity we discussed earlier. In "Towards Identifying Personalized Twitter Trending Topics using the Twitter Client RSS Feeds", *Jinan Fiaidhi*, *Sabah Mohammed*, and *Aminul Islam*, take on the challenge of mining the massive, real-time stream of tweets for interesting trending topics. Moreover, the notion of what is interesting can be personalized for each user based not only on the tweets' vocabulary, but also on the user's personal details and geographical location. Their paper thus defines the first true Twitter stream personalization system.

Staying on the topic of defining innovative new systems, in "Architecture of a Cloud-Based Social Networking News Site", three undergraduate engineering students, *Jeff Luo*, *Jon Kivinen*, and *Joshua Malo*, give us a tour of a social networking platform they developed. Their work presents a new perspective on web data mining from social networks, in which every aspect of the social network is under the control of the researchers, from the type of information users can put up to the underlying cloud architecture itself.

In "Analyzing Temporal Queries for Improving Web Search", *Rim Faiz* brings us back to the topic of web query understanding. Her work focuses on the challenge of adding a temporal understanding component in web search systems. Mining temporal information in this way can help improve search engines by making it possible to correctly interpret queries that are dependent on temporal context. And indeed, her enhanced method shows a promising increase in accuracy.

In "Trend Recalling Algorithm for Automated Online Trading in Stock Market", *Simon Fong*, *Jackie Tai*, and *Pit Pichappan*, exploit another source of web data: the online stock market. This data source is one often overlooked (only 75 references for the entire 2002-2011 period we studied in Figure 1), but one of unquestionable importance today. This paper shows how this data can be mined for trends, which can then be used to successfully guide trading decisions. Specifically, by matching the current trend to past trends and recalling the trading strategies that worked in the past, the system can adapt its behaviour and greatly increase its profits.

In a further example of both the variety of web data and of data mining tasks, in "A Novel Method of Significant Words Identification in Text Summarization", *Maryam Kiabod*, *Mohammad Naderi Dekhordi* and *Mehran Sharafi*, mine a database of web newswire in order to train a neural network to mimic the behaviour of a human reader. This neural network underlies the ability of their system to pick out the important keywords and key sentences that summarize a document. Once trained on the web data set, the system works quite well, and in fact outperforms commercially-available summarization tools.

Our final two papers take a wider view of the challenge of web data mining, and focus on the mining process itself. In our penultimate paper, "Attribute Overlap Minimization and Outlier Elimination as Dimensionality Reduction Techniques for Text Classification Algorithms", *Simon Fong* and *Antonio Cerone* note that the massive and increasing volume of online documents, and the corresponding increase in the number of features to be handled to represent them, is becoming a problem for web mining algorithms, and especially for real-time algorithms. They thus explore the challenge of feature reduction in web documents. Their experiments – conducted on Wikipedia articles in multiple languages and on CNN.com news articles – demonstrate not only the possibility but also the benefits of dimensionality reduction of web data.

Our final paper, "New Metrics between Bodies of Evidence" by *Pascal Djiknavorian*, *Dominic Grenier*, and *Pierre Valin*, presents a higher-level theoretical perspective on web data mining. They propose new metrics to compare and evaluate evidence and uncertainty in the context of the Dempster-Shafer theory. Their work introduces fundamental theoretical advances of consequence for all information retrieval applications. It could be useful, for example, for a new generation of web search systems that can pinpoint relevant information in a web page, rather than consider the page as a whole. It could also be useful to handle the uncertainty incurred when combining information from multiple heterogeneous web data sources.

Richard Khoury received his Bachelor's Degree and his Master's Degree in Electrical and Computer Engineering from Laval University (Québec City, QC) in 2002 and 2004 respectively, and his Doctorate in Electrical and Computer Engineering from the University of Waterloo (Waterloo, ON) in 2007. Since August 2008, he has been an Assistant Professor, tenure track, in the Department of Software Engineering at Lakehead University. Dr. Khoury has published 20 papers in international journals and conferences, and has served on the organization committee of three major conferences. His primary area of research is natural language processing, but his research interests also include data mining, knowledge management, machine learning, and intelligent systems.