Journal of Computer and Communications, 2015, 3, 189-195
Published Online May 2015 in SciRes. http://www.scirp.org/journal/jcc
http://dx.doi.org/10.4236/jcc.2015.35024

Scientific Research Publishing

Big Data Stream Analytics for Near Real-Time Sentiment Analysis

Otto K. M. Cheng, Raymond Lau

Department of Information Systems, City University of Hong Kong, Hong Kong, China
Email: csyklau@yahoo.com

Received April 2015

Abstract

In the era of big data, huge volumes of data are generated from online social networks, sensor networks, mobile devices, and organizations' enterprise systems. This phenomenon provides organizations with unprecedented opportunities to tap into big data to mine valuable business intelligence. However, traditional business analytics methods may not be able to cope with the flood of big data. The main contribution of this paper is the illustration of the development of a novel big data stream analytics framework named BDSASA that leverages a probabilistic language model to analyze the consumer sentiments embedded in hundreds of millions of online consumer reviews. In particular, an inference model is embedded into the classical language modeling framework to enhance the prediction of consumer sentiments. The practical implication of our research work is that organizations can apply our big data stream analytics framework to analyze consumers' product preferences, and hence develop more effective marketing and production strategies.

Keywords

Big Data, Data Stream Analytics, Sentiment Analysis, Online Review

1. Introduction

In the era of the Social Web, user-contributed contents have become the norm. The amounts of data produced by individuals, business, government, and research agents have been undergoing an explosive growth—a phenomenon known as the data deluge. For individual social networking, many online social networking sites have between 100 and 500 million users. By the end of 2013, Facebook and Twitter had 1.23 and 0.64 billion active users, respectively. The number of friendship edges of Facebook is estimated to be over 100 billion. The stream of huge amounts of user-contributed contents, such as online consumer reviews, online news, personal dialogs, search queries, and so on, have called for the research and development of a new generation of analytics methods and tools to effectively process them, preferably in real-time or near real-time. Big data is often characterized by three dimensions, named the 3 V's: Volume, Velocity, and Variety [1]. Currently, there are two common approaches to deal with big data, namely batch-mode big data analytics and streaming-based big data analytics.

Most data originally produced from the Social Web is streaming data. For example, the data representing ac-

How to cite this paper: Cheng, O.K.M. and Lau, R. (2015) Big Data Stream Analytics for Near Real-Time Sentiment Analysis. *Journal of Computer and Communications*, **3**, 189-195. http://dx.doi.org/10.4236/jcc.2015.35024

tions and interactions among individuals in online social media, or the data denoting some events captured by sensor networks is the typical kind of streaming data. Other types of big data perhaps are just a snapshot view of the streaming data generated from a specific point of time. The distinguished characteristic of a big data stream is that data continuously arrive at high speed. Accordingly, effective big data stream analytics methods should process the streaming data in one go, and under very strict constraints of space and time. Currently, research about big data analytics algorithms often focuses on processing big data in batch mode, while algorithms designed to process big data stream in real-time or near real-time are not abundant.

Figure 1 depicts a taxonomy of the common approaches (tools) for processing big data. Big data analytics approaches can be generally divided into distributed or single host approaches. For distributed big analytics methods, there can be then further classified into batch mode processing or streaming mode processing. Even though batch mode big data analytics methods (e.g., MapReduce) are the current dominated method, online incremental algorithms that can effectively process continuous and evolving data stream are desirable to address both the "volume" and the "velocity" issue of big data pasted on online social media. MapReduce and big data stream analytics are two different classes of analytical approaches although they are related for certain theoretical perspectives. Recently, researchers and practitioners have tried to integrate streaming-based analytics and online computation on top of the MapReduce batch mode analytics framework. Sample tools of that kind include the Hadoop Online Prototype. However, more research should be conducted for the development of next generation of big data stream analytics methods that inherit the merits from both batch mode analytics and streaming analytics.

The main contribution of this paper is the design and development of a novel big data stream analytics framework that provides the essential infrastructure to operationalize a probabilistic language modeling approach for near real-time consumer sentiment analysis. There is significant research and practical value of our work because organizations can apply our framework to better leverage the collective social intelligence to develop effective marketing and product design strategies. As a result, these organizations become more competitive in the global marketplace, which is one of the original promises of big data analytics.

With the rapid growth of the Social Web, increasingly more Web users have posted and extracted viewpoints about products, people, or political issues via a variety of online social media such as Blogs, forums, chat-rooms, and social networks. The big volume of user-contributed contents opens the door for automated extraction and analysis of the sentiments or emotions referring to the underlying entities such as consumer products. Sentiment analysis is also referred to as opinion analysis, subjectivity analysis, or opinion mining [2] [3]. Sentiment analysis aims to extract subjective feelings about some subjects rather than simply extracting the objective facets about these subjects [4]. Analyzing the sentiments of messages posted to social networks or online forums can generate countless business values for the organizations which aim to extract timely business intelligence about how their products or services are perceived by their customers [5]. Other possible applications of sentiment analysis include the analysis of the propaganda and activities of cybercriminal groups who pose serious threats to business or government owned web sites [2].

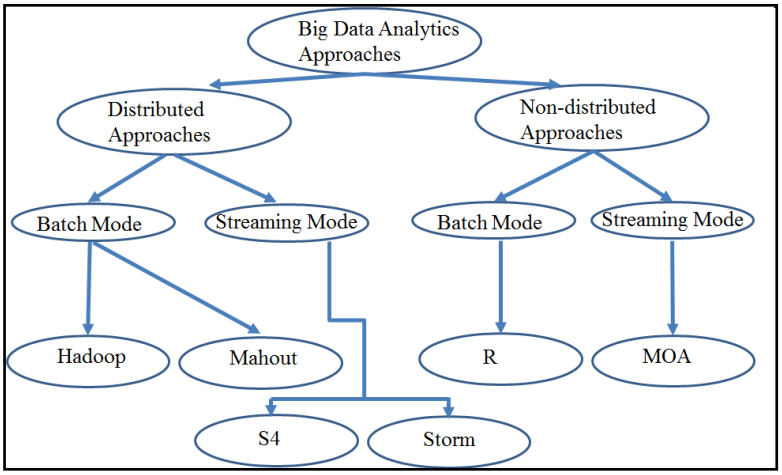

Figure 1. A taxonomy of big data analytics approaches.

Sentiment analysis can be applied to a phrase, a sentence, or an entire message [4]. Most of the existing sentiment analysis methods can be divided into two main camps. The first common paradigm utilizes a sentiment lexicon or heuristic rules as the knowledge base to locate opinionated expressions and predict the polarity of these opinioned expressions [3] [6]. The second common approach of sentiment analysis is based on statistical learning methods [4]. Nevertheless, each camp has its own limitations. For instance, for the lexicon-based methods, common sentiment lexicons may not be able to detect the context-sensitive nature of opinion expressions. For example, while the term "small" may have a negative polarity in a hotel review that refers to a "small" hotel room, the same term could have a positive polarity such as "a small and handy notebook" in consumer reviews about computers. In fact, the token "small" is defined as a negative opinion word in the well-known OpinionFinder sentiment lexicon.

In contrast, statistical learning techniques such as supervised machine learning method usually requires a large number of labeled training cases in order to build an effective classifier to identify the polarity of opinionated expressions. Unfortunately, it is not practical to assume the availability of a large number of human labeled training examples, particularly in a big data environment. On the other hand, both approaches may not be scale up to analyze a huge number of opinioned expressions as found in nowadays Social Web. There is an obvious research gap to develop new methods to be able to analyze big social media data in real-time or near real-time by leveraging a parallel and distributed system architecture. Our research work reported in this paper just tries to fill such a research gap.

The business implication of our research is that business managers and product designers can apply the proposed big data stream analytics framework to more effectively and promptly analyze the consumer sentiments embedded in online consumer reviews. As a result, proactive marketing or product design strategies can be developed to enhance the business operations and the competitive power of the corresponding firms. Moreover, third-party reputation monitoring agencies can apply the proposed framework to continuously monitor the sentiments toward the targeted products and services, and extract appropriate social intelligence from online social media in near real-time.

2. The Big Data Stream Analytics Framework

An overview of the proposed framework that leverages Big Data Stream Analytics for online Sentiment Analysis (BDSASA) is depicted in **Figure 2**. The BDSASA framework consists of seven layers, namely data stream layer, data pre-processing layer, data mining layer, prediction layer, learning and adaptation layer, presentation layer, and storage layer. For these layers, we will apply sophisticated and state-of-the-art techniques for rapid service prototyping. For instance, Storm, the open-source Distributed Data Stream Engine (DDSE) for big data is applied to process streaming data fed from dedicated APIs and crawlers at the Data Stream Layer. For instance, the Topsy API is used to retrieve product related comments from Twitter.

The Storage Layer leverages Apache HBase and HDFS for real-time storage and retrieval of big volume of consumer reviews discussing products and services. The Stanford Dependency Parser and the GATE NER module [7] are applied to build the Data Pre-processing Layer. Our pilot tests show that the size of the multilingual social media data streams is within the range between 0.2 and 0.4 Gigabytes on a daily basis, and this volume is steadily growing. For the feature extraction layer, the Affect Miner utilizes a novel community-based affect intensity measure to predict consumers' moods towards products. Among the big six classes *i.e.*, anger, fear, happiness, sadness, surprise, and neutral commonly used in affect analysis, we focus on the anger, fear, sadness, and happiness classes relevant for product sentiment analysis. The WordNet-Affect lexicon [8] extended by a statistical learning method is used by the Affect Miner. Since social media messages are generally noisy, one novelty of our framework is that we reduce the noise of the "affect intensity" measure by processing messages really related to consumers' comments about products or services.

Previous research employed the HMM method to mine the latent "intents" of actors [9]. We exploit a novel and more sophisticated online generative model and the corresponding distributed Gibbs sampling algorithm to build our Latent Intent Extractor that predicts the intents of consumers for potential product or service acquisitions. The Sentiment Extractor utilizes well-known sentiment lexicons such as OpinionFinder to extract the sentiment words embedded in consumer reviews. Finally, overall sentiment polarity prediction for consumer reviews is performed based on a novel inferential language modeling method. The computational details of this inferential language modeling method for context-sensitive sentiment analysis will be explained in the next section. The overall sentiment polarity against a product or a product category is communicated to the user of the

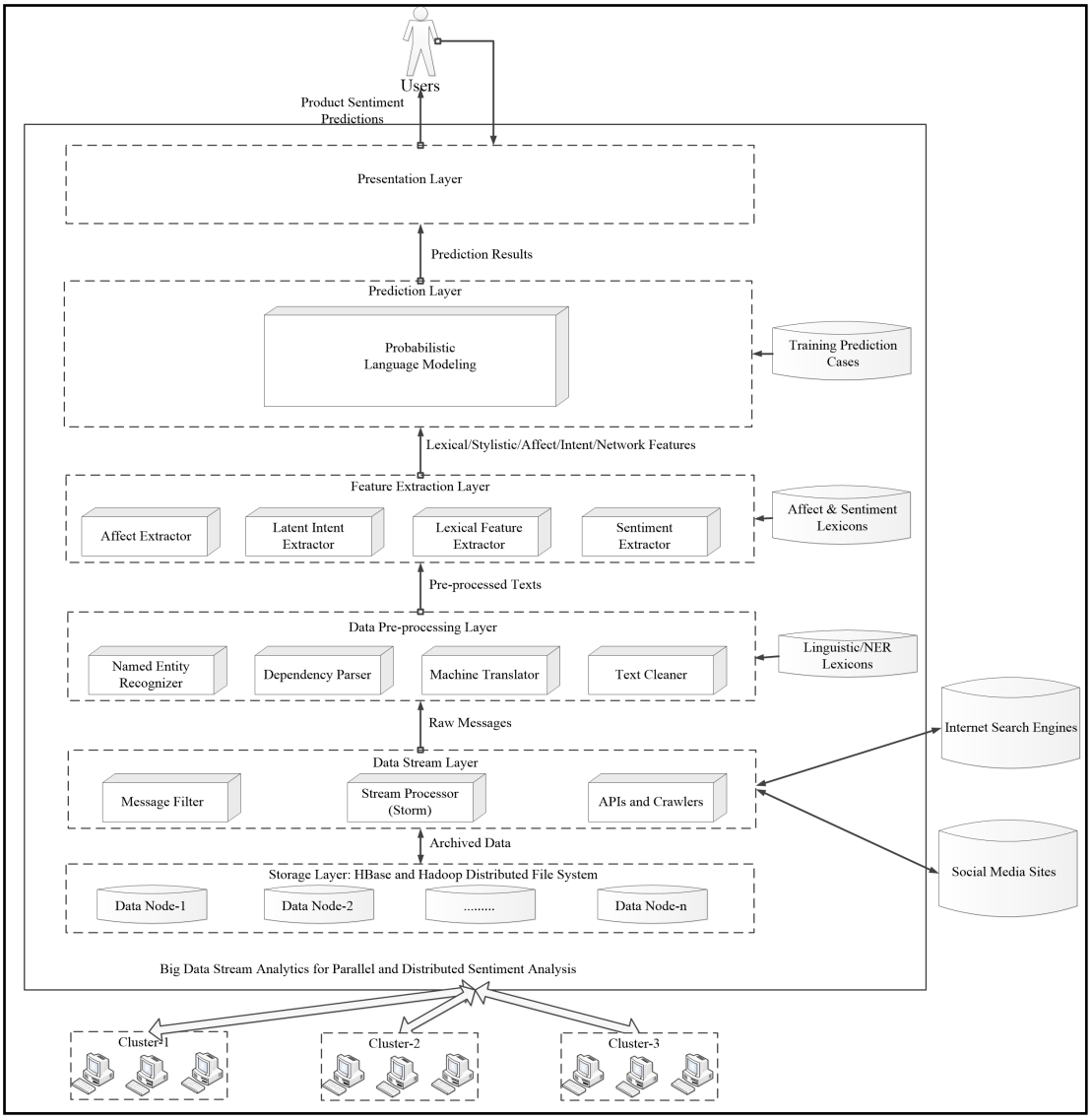

Figure 2. An overview of the BDSASA framework.

system via the presentation layer. Different modes of presentations (e.g., text, graphics, multimedia on desktops or mobile devices) are supported by our framework.

In addition, a novel parallel co-evolutionary genetic algorithm (PCGA) is designed so that the proposed prediction model is equipped with a learning and adaptation mechanism that continuously tunes the whole service with respect to possibly changing features of the problem domain. The PCGA can divide a large search space into some subspaces for a parallel and diversified search, which improves both the efficiency and the effectiveness of the heuristic search process. Each subspace (*i.e.*, a sub-population) is hosted by a separate cluster. Three fundamental decisions are involved for the design a genetic algorithm (GA), that is, a fitness function, chromosome encoding, and a procedure that drives the evolution process of chromosomes [10]. First, the fitness function of our PCGA is developed based on a performance metric (e.g., accuracy of sentiment polarity prediction). Second, since various components of the proposed service should be continuously refined, there are multiple sub-populations of chromosomes to be encoded and co-evolved simultaneously. During each evolution cycle, the best chromosome of a sub-population (e.g., prediction features, social media sources, system parameters) is exchanged with that of other sub-populations. Armed with all the essential information, each chromosome of a sub-population represents a feasible prediction, and its fitness can be assessed accordingly.

3. Probabilistic Language Modeling for Sentiment Analysis

Originally, the term "language model" has been widely explored in the speech recognition community, and it refers to a probability distribution which represents the statistical regularities for the generation of a language [11]. In other words, a language model is a probabilistic function that assigns a probability mass to a string drawn from some vocabulary. In the context of Information Retrieval (IR), a language model M_d is used to estimate the probability that a document d generates a query q [12]. In particular, such a probabilistic inference is used to mimic the concept of document "relevance" of d respect to q. The basic unigram language model is defined according to the following formulas [12] [13]:

$$P(q \mid d) \propto P(q \mid M_d) = \prod_{t \in q} P(t \mid M_d) \tag{1}$$

$$P(t \mid M_d) = (1-\lambda) P_{ML}(t \mid M_d) + \lambda P_{ML}(t \mid M_D) \tag{2}$$

$$P_{ML}(t \mid M_d) = \frac{tf(t,d)}{|d|} \tag{3}$$

where M_d is the language model of the document d. With Jelinek-Mercer smoothing [13], the probability of the document generating a query term t (*i.e.*, $P(t \mid M_d)$) is estimated according to the maximum likelihood model $P_{ML}(t \mid M_d)$, and the maximum likelihood model of the entire collection $P_{ML}(t \mid M_D)$. λ is the Jelinek-Mercer smoothing parameter [13]. The smoothing process is used to alleviate the problem of over-estimating the probabilities for query terms found in a document and the problem of under-estimating the probabilities for terms not found in the document. The function $tf(t,d)$ returns the term frequency of term t in the document d, and $|d|$ is the document length measured by the number of tokens contained in the document.

However, previous studies found that applying the probabilities of query related terms of a relevant context instead of the probabilities of the individual query terms estimated based on the entire document collection (*i.e.*, a general product review context) to a document language model will lead to a more effective smoothing process, and hence lead to good IR performance [14]. Following the similar kind of idea, we develop an inferential language model to compute the probability that a document d (e.g., a product review) will generate a term t found in a Sentiment Lexicon (SL). In order to ensure a more robust and effective smoothing process, the inferential language model can take into account terms (opinion evidences) associated with the opinion indicators in a relevant online review context. In particular, the associated opinion evidences are discovered based on the context-sensitive text mining process over an online review context. The inferential language model for context-sensitive opinion scoring is then defined as follows.

$$P(SL \mid d) \propto P(SL \mid M_d) = \prod_{t \in SL} P(t \mid M_d) \tag{4}$$

$$P(t \mid M_d) = (1-\lambda) P_{ML}(t \mid M_d) + \lambda P_{INF}(t \mid M_d) \tag{5}$$

$$P_{INF}(t \mid M_d) = \tanh\left(\sum_{(t \to t') \in OE} P(t \to t') \cdot P_{ML}(t' \mid M_d)\right) \tag{6}$$

where $P(SL \mid d)$ is the document language model for estimating the probabilities that the document d will generates the opinion indicators defined in a sentiment lexicon (SL). However, to address the common problem that sentiment lexicons may not capture all possible sentiments of a problem domain (e.g., context-sensitive opinion evidences are missing), the proposed language model can take into account other opinion evidences contained in the document by means of the inferential language model $P_{INF}(t \mid M_d)$. The set of context-sensitive opinion evidences OE is dynamically generated according to a context-sensitive text mining technique.

The term association (term inference) of the form $t \to t'$ is applied to the inferential language model to compute the probability that a document generates a term (e.g., an opinion indicator) which is contextually associated with another opinion indicator captured in a sentiment lexicon [15]. For easy of implementation, we only include the top χ term associations captured in OE for each opinion indicator t. It should be noted that the inference that d generating t' involves a certain degree of uncertainty. As a result, the maximum likelihood estimation of $P_{ML}(t' \mid M_d)$ is moderated by a factor $P(t \to t')$. The hyperbolic tangent function is applied to moderate the probability function $P_{INF}(t \mid M_d)$ such that its values fall in the unit interval.

4. Discussions and Summary

While some research work has been devoted to big data analytics recently, very few studies about big data

stream analytics are reported in the literature. The main theoretical contributions of our research include the design and development of a novel big data stream analytics framework, named BDSASA for the near real-time analysis of consumer sentiments. Another main contribution of this paper is the illustration of a probabilistic inferential language model for analyzing the sentiments embedded in an evolving big data stream generated from online social media. The business implication of our research is that business managers and product designers can apply the proposed big data stream analytics framework to more effectively analyze and predict consumers' preferences about products and services. Accordingly, they can take proactive business strategies to streamline the marketing or product design operations.

One limitation of our current work is that the proposed framework has not been tested under an empirical setting. We will devote our future effort to evaluating the effectiveness and efficiency of the BDSASA framework based on realistic consumer reviews and social media messages collected from the Web. On the other hand, we will continue to refine the proposed inferential language model for better sentiment polarity prediction. For instance, a consumer may connect to other consumers via a social network. We may incorporate such connection features in the inferential language model when the sentiment polarity of a review is analyzed. Moreover, the prediction thresholds for probabilistic opinion scoring will be fine-tuned using the proposed PCGA. Finally, we will conduct a usability study for the proposed big data stream analytics service in a real-world e-Business environment.

Acknowledgements

This research work was partially supported by a grant from the Shenzhen Municipal Science and Technology R & D Funding—Basic Research Program (project number: JCYJ20130401145617281 and project number: JCYJ20140419115614350).

References

[1] Boden, C., Karnstedt, M., Fernandez, M. and Markl, V. (2013) Large-Scale Social-Media Analytics on Stratosphere. *Proceedings of the* 22*nd International Conference on World Wide Web Companion*, 257-260.

[2] Lau, R.Y.K., Xia, Y. and Ye, Y. (2014) A Probabilistic Generative Model for Mining Cybercriminal Networks from Online Social Media. *IEEE Computational Intelligence Magazine*, **9**, 31-43. http://dx.doi.org/10.1109/MCI.2013.2291689

[3] Turney, P.D. and Littman, M.L. (2003) Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM Transactions on Information Systems*, **21**, 315-346. http://dx.doi.org/10.1145/944012.944013

[4] Wilson, T., Wiebe, J. and Rwa, R. (2004) Just How Mad Are You? Finding Strong and Weak Opinion Clauses. In: McGuinness, D.L. and Ferguson, G., Eds., *Proceedings of the Nineteenth National Conference on Artificial Intelligence, Sixteenth Conference on Innovative Applications of Artificial Intelligence*, San Jose, 25-29 July 2004, 761-769.

[5] Archak, N., Ghose, A. and Ipeirotis, P.G. (2007) Show Me the Money!: Deriving the Pricing Power of Product Features by Mining Consumer Reviews. In: Berkhin, P., Caruana, R. and Wu, X., Eds., *Proceedings of the* 13*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, 12-15 August 2007, 56-65. http://dx.doi.org/10.1145/1281192.1281202

[6] Turney, P.D. (2002) Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the* 40*th Annual Meeting of the Association for Computational Linguistics*, 417-424.

[7] Maynard, D., Tablan, V., Ursu, C., Cunningham, H. and Wilks, Y. (2001) Named Entity Recognition from Diverse Text Types. *Proceedings of the* 2001 *Conference on Recent Advances in Natural Language Processing*, Tzigov Chark, Bulgaria.

[8] Valitutti, A., Strapparava, C. and Stock, O. (2004) Developing Affective Lexical Resources. *Psychology*, **2**, 61-83.

[9] Zhang, Q., Man, D. and Wu, Y. (2009) Using HMM for Intent Recognition in Cyber Security Situation Awareness. *Proceedings of the Second IEEE International Symposium on Knowledge Acquisition and Modeling*, 166-169. http://dx.doi.org/10.1109/kam.2009.315

[10] Lau, R.Y.K., Tang, M., Wong, O., Milliner, S. and Chen, Y. (2006) An Evolutionary Learning Approach for Adaptive Negotiation Agents. *International Journal of Intelligent Systems*, **21**, 41-72. http://dx.doi.org/10.1002/int.20120

[11] Nadas, A. (1984) Estimation of Probabilities in the Language Model of the IBM Speech Recognition System. *IEEE Transactions on Acoustics, Speech and Signal Processing*, **32**, 859. http://dx.doi.org/10.1109/TASSP.1984.1164378

[12] Ponte, J.M. and Croft, W.B. (1998) A Language Modeling Approach to Information Retrieval. *Proceedings of the* 21*st*

Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 275-281. http://dx.doi.org/10.1145/290941.291008

[13] Zhai, C.X. and Lafferty, J. (2004) A Study of Smoothing Methods for Language Models Applied to Information Retrieval. *ACM Transactions on Information Systems*, **22**, 179-214. http://dx.doi.org/10.1145/984321.984322

[14] Nie, J.-Y., Cao, G.H. and Bai, J. (2006) Inferential Language Models for Information Retrieval. *ACM Transactions on Asian Language Information Processing*, **5**, 296-322. http://dx.doi.org/10.1145/1236181.1236183

[15] Lau, R.Y.K., Song, D., Li, Y., Cheung, C.H. and Hao, J.X. (2009) Towards a Fuzzy Domain Ontology Extraction Method for Adaptive E-Learning. *IEEE Transactions on Knowledge and Data Engineering*, **21**, 800-813. http://dx.doi.org/10.1109/TKDE.2008.137