Mining the blogosphere for contributor’s sentiments

Boris Galitsky

School of Computer Science and Information Systems
Birkbeck College, University of London
Malet Street, London WC1E 7HX, UK
galitsky@dcs.bbk.ac.uk

Boris Kovalerchuk

Dept. of Computer Science,
Central Washington University,
Ellensburg, WA, 98926, USA
borisk@cwu.edu

Abstract

We propose a domain-independent approach to blog analysis for an assessment of bloggers’ sentiments. The generic methodology to detect specific attitudes of human agents is based on representation of conflict scenarios by means of communicative actions (which express sentiments) and causal/argumentative links between their subjects (which may support these sentiments). This methodology is applied to textual data on inter-human interactions including conflicts in opinions of bloggers. Evaluation is conducted in the domain of customer complaints.

Introduction

Weblogs can be defined as a website that contains an online personal journal with ordered entries, reflections, comments, and hyperlinks provided by a writer. Weblogs provide unprocessed highly biased personal comments from a wide audience of contributors. Usually weblogs are ordered sequences of entries with hyperlinks to other resources. Collections of weblogs can play such roles as tracking the public opinion, retaining the communities which exist online, facilitate online collaboration and others. Analysis and mining of weblogs is becoming an important activity nowadays (Glance et al 2005).

Weblogs are frequently viewed as an upcoming substitute for mass media. Readership is increased by 58% in 2004; 7% of Internet users are blogging (8 million Americans), and 27% of Internet users read weblogs (32 million Americans)• 38% of Internet users know what a weblog is (46 million Americans), in accordance to (BlogPulse 2005).

Analyzing weblogs, it is quite straightforward to categorize them with respect to a domain of activity (travel, political, consumer etc.) using keyword-based statistical methods (Slattery &Craven 1998; Intelliseek 2005). However, a deeper-level analysis of weblogs is rather hard due to a lack of structure, style, logical sequence and thoroughness of references as reported by (Gorman 2005).

What kind of domain-independent analysis can be applied to the weblogs to characterize the bloggers? In this paper we propose the framework to mine for the sentiments of bloggers, extracting the patterns of their interactions with each other, as mentioned in blog text. These patterns are sought to be the main characteristics of the communities of bloggers (blogosphere), and represent valuable information for marketing, cultural and social studies, as well as security applications.

Scenarios of interaction between agents are an important subject of study in AI. Weblogs are an extremely valuable source of information on inter-human conflicts, because news and information are posted immediately in the form of (interactive) comments which are rather direct and sincere assessment of knowledge about a domain and involved agents. The amount of such data is practically unlimited for any kind of computational analysis.

There is another important role of the assessment of bloggers’ mood. Detecting the authors with bad mood and inappropriate argument patterns, it is possible to apply the access control and automated filtering to improve the content and/or enforce certain blogs’ policy. Hence we believe an accurate assessment of contributors’ mood is essential in a weblog operation and should be accepted as its necessary component in the near future.

Describing behavior of bloggers, we restrict our considerations to communicative actions (and causal links between them) of individual bloggers and their communities in the course of interaction (conflict). Behavior of real-world conflicting agents in weblogs is described in a richer language using a wide number of mental entities including *pretending*, *deceiving*, *offending*, *forgiving*, *trust*, etc. In this paper we take advantage of our earlier result that following the logical structure of how negotiations are represented in a scenario (text or structure), it is possible to judge on the consistency of this scenario (Galitsky & Tumarkina 2004).

Expressing sentiments in weblogs

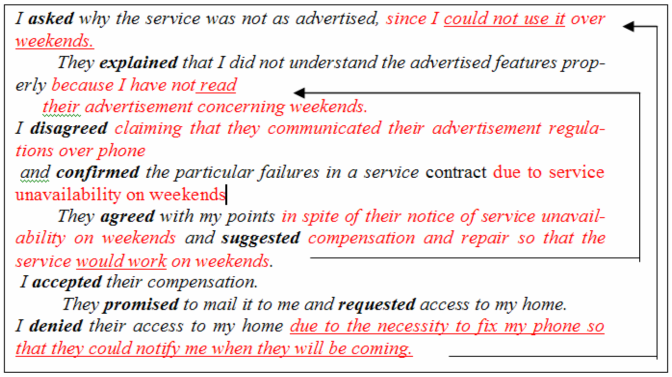
In this section we present our model of a scenario in a weblog. A scenario is extracted from a single posting as a text or from a few postings as both text and submission sequence. Here we develop a knowledge representation methodology based on approximation of a natural language description of a conflict (Galitsky 2003).

To form a data structure for machine learning, we represent a weblog as a sequence of communicative actions, ordered in time, with a causal relation between certain communicative actions (Fig.1). Scenarios are simplified to allow for effective matching by means of graphs. Only communicative actions remain as the most important component to express similarities between scenarios. Each vertex corresponds to a communicative action, which is performed by either *proponent*, or *opponent*, the latter are called *agents*. An arc (oriented edge) denotes a sequence of two actions.

In our model mental actions have two parameters: *agent name* and *subject* (information transmitted, a cause addressed, a reason explained, an object described, etc.). Representing scenarios as graphs, we take into account both parameters. Arc types bear information whether the subject stays the same. Thick arcs link vertices that correspond to communicative actions with the same subject; thin arcs link vertices that correspond to communicative actions with different subject. The curve arcs denote a causal or argumentative link between the arguments of mental actions.

One of the most important tasks in assisting negotiations and resolving inter-human conflicts is assessment of *sentiments* of involved parties. An agent's mood displayed in a single posting or a sequence of postings is *positive* if the proposition is plausible, properly communicated, backed up by appropriate argumentation, internally consistent, and also consistent with available domain-specific knowledge, and *negative* otherwise. In case of inter-human discussions or negotiations, such domain-specific knowledge is frequently unavailable. In most cases it is hard to build ontology or formulate explicit rules for positive and negative sentiments. Instead, a given posting is positive (negative) if it is similar to an earlier posting which is considered to be positive (negative, respectively) by an expert.

In this study we explore the possibility to relate a weblog message to a class of *positive* and *negative* attitudes of involved human agent (including the author of the message). These attitudes are specified by experts using examples (a training dataset); we do not envision a possibility to express attitudes via explicit rules.



*I **asked** why the service was not as advertised, since I could not use it over weekends.*

*They **explained** that I did not understand the advertised features properly because I have not read their advertisement concerning weekends.*

*I **disagreed** claiming that they communicated their advertisement regulations over phone*

*and **confirmed** the particular failures in a service* contract due to service unavailability on weekends

*They **agreed** with my points in spite of their notice of service unavailability on weekends and **suggested** compensation and repair so that the service would work on weekends.*

*I **accepted** their compensation.*

*They **promised** to mail it to me and **requested** access to my home.*

*I **denied** their access to my home due to the necessity to fix my phone so that they could notify me when they will be coming.*

Fig. 1a. A scenario which includes communicative actions of a proponent and an opponent, and argumentative attack relation between the subjects of these actions.

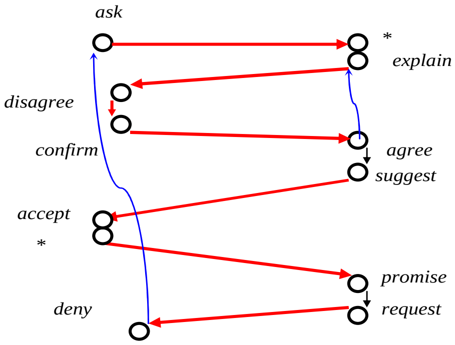


Fig. 1b. The graph-based representation for blog scenario Fig. 1a.

Analyzing consumer advocacy blogs

In this section we conduct the evaluation of the customer complaint weblog. The textual complaints we used were downloaded from the public website PlanetFeedback.com during 3 months starting from March 2004. Our training dataset contains textual complaints similar to those in a number of weblogs like (BuzzMachine 2005; TheLoyaltyLeader 2005). For the purpose of this evaluation, each complaint was manually coded as a sequence of communicative actions and assigned a status by us.

We used the Nearest Neighbor technique (Mitchell, 1997) for relating a blog graph to a class of *positive* (adequate) or *negative* (inadequate) blogs, having defined the *operation of similarity* as intersection between graphs obeying certain properties (Galitsky et al 2005).

We used the data for fourteen companies, 20 complaints for training and 20 complaints for evaluation for each company. Firstly, the consistency of each training dataset is evaluated when blogger's attitude for each complaint is assumed *unknown* and classification is performed. After that the numbers of false positives, false negatives, and correct classification results are obtained.

The resultant recognition accuracy is 70.4%. Being quite low in accordance to pattern recognition standards in such domains as speech and visual object recognition, this accuracy is believed to be satisfactory for the weblog mining problem setting.

Related work and conclusions

We mention a number of weblogs mining tools such as (Blogarama 2005, Blogstreet 2005 and Syndic8 2005). Machine learning has been heavily deployed for the analysis of distributed resources like weblogs (Kleinberg 1999; Chakrabarti et al 1998; Glance at al 2005, Kohavi et al 2004), including inductive learning approaches (Slattery & Craven 1998; Craven 1998).

Weblogs are currently subject to such processing methods as monitoring weblogs, wrapping weblogs to extract posts, indexing and search over weblogs, trend mining including search for top links, key phrases, key people, weblogbites, and trend search. Aggregation of weblogs is an important means which assists in extraction of public opinions from weblogs. The web-based aggregators, including Bloglines, My Yahoo!, Feedster, Newsgator and desktop aggregators such as Pluck (Internet Explorer plug-in) and Sage (Firefox plug-in) play an important role. However, to the best of our knowledge, no information extraction concerning attitudes of contributing bloggers is currently conducted.

A number of studies in game research and distributed AI communities have been addressing the problem of learning communicative actions and multiagent interactions (e.g. Kalai and Lehrer 1993). However, to mine for agents' sentiments from text, it is necessary to improve the competence in expressing agents' attitudes. Learning in inter-human setting is closely related to modeling the overall interactions process in terms of communicative actions, i.e., what negotiation protocols are adopted.

We proposed a machine learning-based approach to relate a formalized scenario, extracted from a weblog posting, to a class. It has been developed and evaluated in the consumer advocacy blogosphere by means of classification of bloggers' moods. The number and structure of classes depend on a domain, but the criteria of sequences of communicative actions have been shown useful to express commonalities between scenarios. The approach to mine blogs as scenarios of inter-human interactions (encoded as sequences of communicative actions) is believed to be original on one hand and universal on the other hand.

Using the proposed analysis, a weblog can be characterized in terms of collective attitude, argumentation merits, and even overall competence of the community. In our further studies we plan to improve the natural language information extraction component which is oriented on communicative actions and their parameters to be obtained from weblogs. Also, analysis of blogger's sentiments would benefit from using visualization techniques (Kovalerchuk and Vityaev 2000, Kovalerchuk & Schwing 2005), heavily used in data mining.

References

BlogPulse.com. (Last accessed Sept 30, 2005).

Blogarama www.blogarama.com (Last accessed Dec 30, 2005)

Galitsky, B., Kuznetsov, S. and Samokhin, M. 2005. Analyzing Conflicts with Concept-Based Learning. Intl. Conf on Concept Structures.

Galitsky, B. 2003. Natural Language Question Answering System Technique of Semantic Headers. *Advanced Knowledge International*, Adelaide, Australia.

Galitsky, B. and Tumarkina, I. 2004. Justification of Customer Complaints using Emotional States and Mental Actions. *FLAIRS* Miami, FL.

Kalai, E., and Lehrer, E. 1993. Rational learning leads to Nash equilibrium, Econometrica 61(5)1019-1045.

Mitchell, T. 1997. Machine Learning, McGraw-Hill.

Osborne, M. J., and Rubinstein, A. A 1994. Course in Game Theory. The MIT Press.

Mellor, K. 2005. Library Stuff 2005 Mining Weblogs for Information
http://www.lori.ri.gov/webdesign/syndication/weblogs.pdf

Intelliseek 2005. http://www.blogpulse.com/papers/

Gorman, M. 2005. Revenge of the Blog People! Library Journal, Feb. 15, 2005
http://www.libraryjournal.com/article/CA502009

Kovalerchuk, B., Vityaev, E. 2000. Data Mining in Finance: Advances in relational and hybrid methods. Kluwer, Boston.

Kovalerchuk, B., Schwing J., (eds) 2005. Visual and Spatial Analysis: Advances in Data Mining, Reasoning, and Problem solving, Springer.

Glance, N., Hurst, M., Nigam, K., Siegler, M., Stockton , R. and Tomokiyo, T. 2005. Analyzing Online Discussion for Marketing Intelligence. Intelliseek Applied Research Center www.intelliseek.com (last accessed Sept 30, 2005).

Slattery, S. and M. Craven 1998. Combining Statistical and Relational Methods for Learning in Hypertext Domains. Proceedings of ILP-98.

Syndic8 www.syndic8.com (Last accessed Dec 30, 2005).

Craven, M. 1998. First Order Learning for Web Mining. European Conference on Machine Learning.

Kleinberg, J.M. 1999. Hubs, Communities and Authorities. Cornell University.

Chakrabarti, S., B. Dom, D. Gibson, J. Keinberg, P. Raghavan, & S. Rajagopalan 1998. Automatic Resource list Compilation by Analyzing Hyperlink Structure and Associated Text. Proc 7th WWW.

Kohavi, R., Mason, L., Parekh, R., Zheng, Z. 2004. Lessons and Challenges from Mining retail e-commerce data. Machine Learning Journal, Special Issue on Data Mining Lessons Learned.

TheLoyaltyLeader 2005.
http://theloyaltyleader.blogspot.com.

BuzzMachine 2005.
http://www.buzzmachine.com/index.php/2005/08/25/customer-service-in-reverse/.