WEBLOG RECOMMENDATION USING ASSOCIATION RULES

Juan Julián Merelo Guervós, Pedro A. Castillo, Beatriz Prieto Campos
Departamento de Arquitectura y Tecnología de Computadores
Universidad de Granada

José Carpio Cañada
Departamento de Electrónica, Sistemas Informáticos y Automática
Universidad de Huelva
Escuela Politécnica Superior

Fernando Tricas García
Departamento de Informática e Ingeniería de sistemas
Universidad de Zaragoza

Gemma Ferreres
Tintachina.com

ABSTRACT

Weblogs are web sites where one or several authors publish their opinions about current events. Even in Spain, there are several thousands, and it is often difficult to find a weblog that meets one's interest. Recommendation services thus become, if not a need, at least a convenience. In this paper we propose automatic extraction of association rules from the results of a survey as a means to recommend to weblog readers other weblogs about related topics, based on the results of a survey. We examine different support/confidence thresholds, and study resulting rules; as a result, we find novel relationships between weblogs, that would not have been found through other means.

KEYWORDS

Weblog, cyberculture, computer communications, virtual communities, media, social relationships, web sites.

1. INTRODUCTION

Information and services offered by the Internet and the World Wide Web (WWW) have grown very fast over the last few years. WWW has become an indispensable medium for up to one billion users around the world.

Among all communication services that Internet offers, one of the fastest growing are weblogs (abbreviated *blogs*) (Blood, 2000; Winer, 1999). Weblogs are web sites where one or more authors publish their opinions about current events, offer comments about other sites or others authors´ opinions. These sites also offer high reader-interaction degree, allowing him/her to publish their comments about original author's opinions.

Some of most popular Spanish blog hosting sites, such as Blogalia (http://www.blogalia.com), contain tens thousands of stories and hundreds of thousands of comments. Browsing among so much information is no easy task, even more when blogs are often updated several times a day, and generic search engines (Google, Yahoo, Altavista, etc) do not update their indices so frequently. Another drawback of current search engines is that they perform only string searches; they not include any semantic information. For example, a searching "Granada" (which, in Spanish, is the city of Granada, in Spain, several other Granada cities all over

the world, and a pomegranate, and a grenade) will give a list of web pages about tourist information, explosives or fruits. That means, is difficult to find which weblogs is talking about a particular topic one is interested in. Disadvantages like that, motivate study of new technologies to generate better results in web knowledge extraction (*web mining*), specifically from weblogs.

This paper is based on applying automatic association rules extraction algorithm. That data mining algorithm, help us to find weblogs not evident associations. We use survey results made among thousands of spanish weblog readers/writers; so far, it is reading habits most extensive survey. Other results, and comments, can be found at http://tintachina.com/archivo/cat_i_encuesta_webloggers.php. By using weblog reading preferences, instead of hyperlinks, our algorithm extracts information from user themselves, who make associations by choosing a few weblogs out of the huge amount of them available. This also means that rules extracted should be expressed as "Readers who chose this (these) weblog(s) also chose this other". It is often the case that weblogs that appear in the same rule, but that does not happen always, for several possible reason: not all weblogs use hyperlinks, and hyperlinks are unidirectional, so hyperlinked need not be aware of that; even more so, those who read a blog often do not know other blogs that link to that one, even as that fact implies that they are related.

This work is organized as follows: after a brief exposition of state of the art, *a priori* algorithm is briefly explained in section 3; further on, a formal problem description and data mining phases is given. Finally, results are presented along with conclusion details and future works.

2. STATE OF THE ART

Association rules are widely used, and can be considered a mainstream procedure for recommendations. Amazon.com, through extensive analysis of clickstreams (visits through their pages), buying habits, as well as other data available for them (wish lists, lists from *listmania*) has one of the most extensive databases of recommendations in the industry, so that readers recommendations are quite accurate. Many other e-commerce sites use same kind of algorithms, but, for the time being, website recommendation is much more elusive.

Main problem is that, even as Amazon's items database is huge, websites number, and even of weblogs, is almost as big. Besides, Amazon features static item with known characteristics, while websites evolve, offer different facets, and change completely, even more so in the case of weblogs, which can be said to be created collaboratively by the author and its readers through comments.

Weblogs have several mechanisms that allow readers to discover others: one of them is simply hyperlinks; if weblogs author finds something interesting elsewhere, he refers to it by adding comments with hyperlink in its own weblog. Another mechanism is called *trackback*: a weblog can inform another that he is linking to it, or talking about a similar topic, by calling a function in a particular way; trackbacks appear alongside commentaries in the post they refer to.

These mechanisms, however, are often not enough. Every weblog works in its own way: some of them do not link at all (think about literary weblogs, for instance, or pure personal diaries), others link more often to news sites than to other weblogs (mainly collective weblogs), and others, simply, are not aware of the huge amount of sites available out there.

So far, we have proposed using community detection mechanisms (Merelo et al., 2004) that help authors and readers to know which community do they belong, and, by linking, what kind of communities emmerge; however, this only tells a part of the story. By using survey result, and applying rule extraction algorithm, we can find new associations that could not have been discovered by analysing links.

3. A PRIORI ALGORITHM

This algorithm was proposed by Rakesh Agrawal and Ramakrishnan Srikant, from the IBM Almaden center (Agrawal et al., 1993, 1994). The problem of discovering all association rules can be decomposed in two different subproblems:

1. Find all itemsets whose support is over a threshold. The support for an itemset is the set of transactions (or *baskets*) that contain that itemset. All itemsets with a support over a threshold are called *large itemsets* and the rest *small itemset*. Algorithms for discovering large itemsets make multiple passes over the data. First pass, we count the support of individual items and determine

which of them are large. In each subsequent pass, we start with a seed set of itemsets found to be large in the previous pass. We use this seed set for generating new potentially large itemsets, called candidate itemsets, and count the actual support for these candidate itemsets during the pass over the data. At the end of the pass, we determine which of the candidate itemsets are actually large, and they become the seed for the next pass. This process continues until no new large itemsets are found.

2. Use large itemsets to generate rules. The general idea is that, if ABCD and AB are large itemsets, then we can determine if the rule AB => CD keeps the ratio *conf* = supp(ABCD)/support(AB). If *conf* ≥ *minconf,* then the rule is valid; ABCD as a rule already has a minimum support since it is a large itemset. To generate rules, for every large itemset *l*, we find all non-empty subsets of *l*. For every such subset *a*, we output a rule of the form *a* => (*l*-*a*) if the ratio of support (*l*) to support(a) is at least *minconf*. We consider all subsets of l to generate rules with multiple consequents. Since the large itemsets are stored in hash tables, the support counts for the subset itemsets can be found effiently.

4. RESULTS

The study has been done on a poll taken among weblog readers and authors, who answered, among others, the following question: ¿Which are the last 3 weblogs you have read?

Data was stored in XML file; after extraction, it had to be cleansed. Some cases, weblog titles or names were used instead of URLs; and even if addresses were used, they sometimes missed the final slash (/) or the initial http://www. At the end of this preprocessing phase, we obtained *baskets* with 3 URLs, with each weblog represented by a single URL. Apriori algorithm described above was then applied to this set of items, and results analyzed.

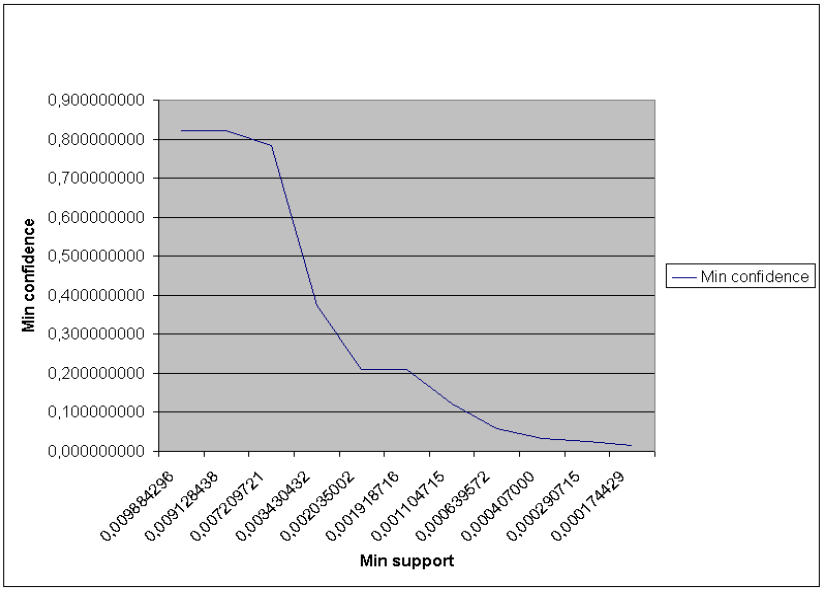

Figure 1. Min confidence vs Min support

Rules were also analyzed to find which was the minimum confidence and support that guaranteed reasonable results. Since even the most popular site (MiniD) was voted only by a fraction of the pollsters, a very small support should be expected; confidence went up with the minimum support, and finally, a level of confidence of around 50%, for a support of around 0.2%, was found reasonable, as is shown in figure 1.

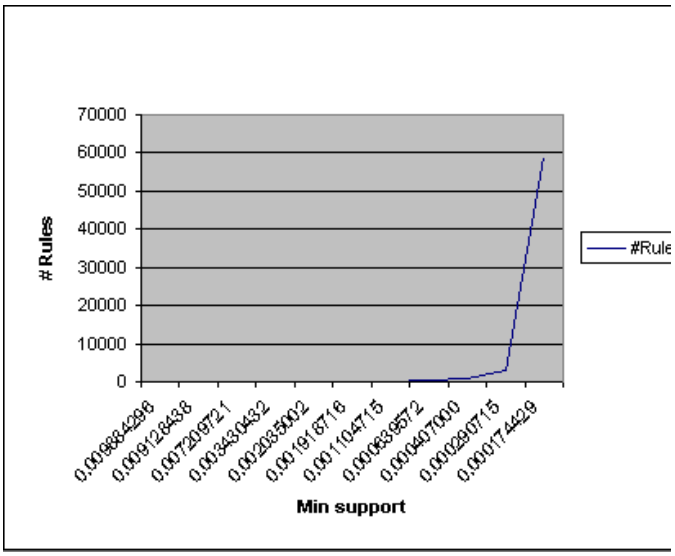

Figure 2. Number of rules changing minimum support

Obviously, the number of rules found changed with confidence level and support chosen, as is shown in figure 2; there were very few rules with a support over 20%, which is to be expected, due to the dispersion of the readership. The confidence of the rules which was finally used in this study was actually quite low, as is shown in figure 3, but due to the huge number of weblogs available, and the number of weblogs that were mentioned in this study (around a few thousands), the fact that two weblogs are associated even in a few polls is quite relevant.

After a detailed analysis of results using our knowledge about the blogosphere and of many of the weblogs used in this study, we have been able to find the relationship among weblogs appearing in the association rules. For example, the following rule:

```
bandaancha.st ==> barrapunto.com (0.692308, 9)
```

with a confidence factor of 0.69 and support of 9 rules represents two collective weblogs with a big audience (in fact, Barrapunto, at http://barrapunto.com, is possibly the most popular blog of the Spanish blogosphere, as was also found in survey results shown in http://tintachina.com/archivo/los_weblogs_mas_leidos.php) and with a strongly technical, and more weakly political, background. In this case, there are several links between both sites, and the result is not altogether unexpected. However, in the following example:

Users who read **barrapunto.com** also read:

1. aretzo.blog-city.com
2. yildelen.blogalia.com
3. laflecha.net
4. osnews.com
5. bulma.net

For this kind of recommendations expert advise would be needed, since they are not so obvious as the previous one. In fact, paper's authors, who are heavy participants in Barrapunto, did not even know about the first one. Using this association rules we can recommend similar weblogs to readers of a given site, without human intervention.

Several groups of tests have been done, changing minimum support and minimum confidence. Minimum support has been reduced until 0,001 in order to obtain 96 rules with minimum confidence of 0,6. For a minimum support of 0,1 no rule is obtained. With a minimum support of 0.0006, 5800 rules were obtained.

From these results we would like to remark the values obtained for rules support. It is very low, but we think they are relevant enough. We think there are two main reasons. The first one is the wide variety of "products" in our "supermarket" which, following example proposed in (Kleinberg, 1999), is very big. Our articles are URLs pointing to weblogs. We have counted more than 2160 articles in this example and we are working with a database of 1473 baskets. We think this is the main reason for a so low min support: great diversity of articles and low number of baskets.

Following figures show number of rules variation with respect to the minimum support and minimum confidence.

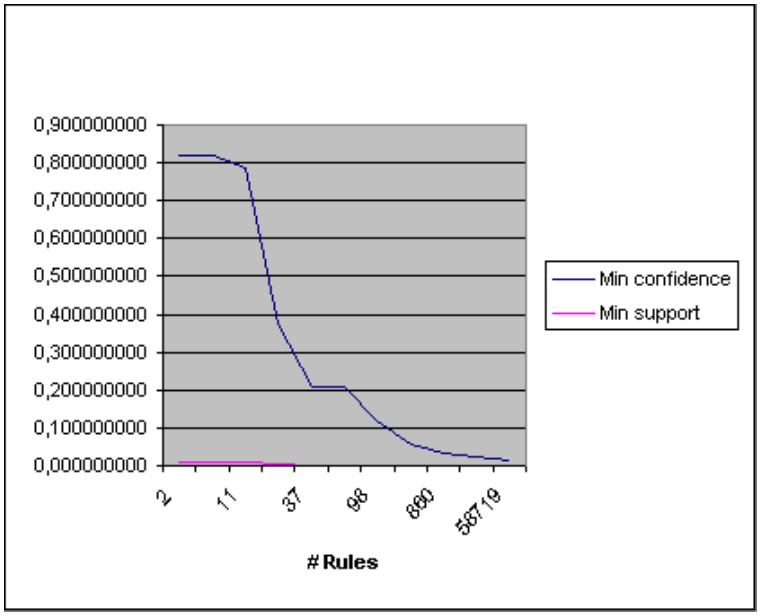

Figure 3. Number of rules changing minimum support and minimum confidence

5. CONCLUSIONS AND FUTURE WORK

We can conclude that obtained results are good enough. Obtained rules offer interesting information for weblogs classification, using data obtained by mining readers replies to a poll. These results can be improved increasing the number of replies. This also makes us think that the "Apriori" algorithm not only works well with big databases, as originally proposed (Agrawal et al., 1993, 1994), but also that it works with smaller ones.

We think this algorithm can be applied also to provide interesting links related to a given post of a weblog. The plan is to use posts as baskets and permalinks, that is, permanent links to weblog postings, as articles.

Another idea is to relate weblogs using URLs in a given weblog as a basket and the weblogs referenced as the set of articles.

Another line of work will be the application of algorithms for obtaining more efficient rules, and also to study methods for obtaining the better min confidence and min support for obtaining optimal rules set.

We also believe that an interesting work on knowledge representation has to be done in order to make these results more useful to weblog users. For example, it would be useful to present this knowledge on-line, together with any given post, in order the reader can access to related content.

ACKNOWLEDGEMENTS

This paper has been funded in part by project TIC2003-09481-C04-04, of the Spanish ministry of science and technology, and a project awarded the Quality and Innovation department of the University of Granada. Fernando Tricas is with the Group of Discrete Event Systems Engineering (GISED), and his work is partially supported by TIC2001-1819, financed by the Spanish Ministerio de Ciencia y Tecnologia. We are also grateful to Antonio Cambronero, of Blogpocket, for his technical support during the realization of the survey.

REFERENCES

Rebecca Blood, "weblogs: a history and perspective", September 2000. Available at www.rebeccablood.net

Dave Winer ,"The History of Weblogs", 1999. Available at newhome.weblogs.com

Juan J. Merelo-Guervós, Beatriz Prieto, Alberto Prieto, Gustavo Romero, Pedro Castillo-Valdivieso, and Fernando Tricas. Clustering web-based communities using self-organizing maps. ,In Kommers et al. First conference on Web Based communities, WBC2004, 158-165 available from http://geneura.ugr.es/jmerelo/papers/72.pdf, Lisbon, March 2004.

R. Agrawal, R. Srikant. "Fast Algorithms for Mining Association Rules", Proc. 1994 Int. Conf. Very Large Database (VLDB'94), pag. 487-499, Santiago, Chile

R. Agrawal, T. Imielinski, A. Swami, "Mining association rules between sets of items in large databases". Proc, ACM SIGMOD, Conf. on Management of Data, pag. 207-216, Washington, D.C., May 1993

J. M. Kleinberg. Authoritative sources in a hyperlinked environment. Journal of ACM, 46 :604-632, 1999

S. Chakrabarti, B.E. Dom, P. Indik. "Enhaced hypertext classification using hyper-links". Proc. 1998 ACM-SIGMOD Int. Conf. Management of Data (SIG-MOD'98) 307-318, Seatle WA, Junio 1998

Minos N. Garofalakis, Rajeev Rastogi, S. Seshadri, Kyuseok Shim, "Data mining and the Web: Past, Present and Future"

Jiawei Han, Micheline Kamber, "Data Mining, Concepts and Techniques", 2001

K. Wang, Szhou, S.C. Liew, "Building hierarchical classifiers using class proximity". Proc. 1999 Int. Conf. Very Large Database (VLDB'99) pag. 363-374, Edimburgo, UK, sept. 1999