FCA-based Recommender Models and Data Analysis for Crowdsourcing Platform Witology *

Dmitry I. Ignatov[1], Alexandra Yu. Kaminskaya[1], Natalia Konstantinova[4], Alexander Malioukov[2], and Jonas Poelmans[1,3]

[1] National Research University Higher School of Economics, Russia, Moscow,
dignatov@hse.ru
http://www.hse.ru
[2] Witology
http://www.witology.com/en
[3] KU Leuven, Belgium
[4] University of Wolverhampton, UK

Abstract. This paper considers a recommender part of the data analysis system for the collaborative platform Witology. It was developed by the joint research team of the National Research University Higher School of Economics and the Witology company. This recommender system is able to recommend ideas, like-minded users and antagonists at the respective phases of a crowdsourcing project. All the recommender methods were tested in the experiments with real datasets of the Witology company.

Keywords: collaborative and crowdsourcing platforms, data mining, Formal Concept Analysis, biclustering, recommender systems

1 Introduction

The success of modern collaborative technologies is marked by the appearance of many novel platforms for distributed brainstorming or carrying out so called "public examination". There are a lot of such crowdsourcing companies in the USA (Spigit[5], BrightIdea[6], InnoCentive[7] etc.) and Europe (Imaginatik[8]). There is also a Kaggle platform[9] which is very beneficial for data practitioners and companies that want to select the best solutions for their data mining problems. In 2011 Russian companies decided to launch business in data mining area as well. Witology[10] and Wikivote[11] are the two most representative examples of

* The final publication is available at http://link.springer.com/chapter/10.1007%2F978-3-319-08389-6_24

[5] http://spigit.com/
[6] www.brightidea.com/
[7] http://www.innocentive.com/
[8] http://www.imaginatik.com/
[9] http://www.kaggle.com
[10] http://witology.com/
[11] http://www.wikivote.ru/

such Russian companies. Several all-Russian projects have already been finished successfully (for example, Sberbank-21[12], National Entrepreneurial Initiative-2012[13] etc.). Socio-semantic networks constitute the core of such crowdsourcing systems [1,2] and new approaches are needed to analyze data underlying these networks.

The term "crowdsourcing" is a portmanteau of "crowd" and "outsourcing", coined by Jeff Howe in 2006 [3]. There is no general definition of crowdsourcing, however it has a set of specific features. Crowdsourcing is a process, both online and offline, that includes a task solving by a distributed large group of people who usually come from different organisations, and are not necessarily paid for their work. As a rule, while participating in a project, users of crowdsourcing platforms discuss and solve one common problem, propose possible solutions and evaluate ideas of each other as experts [3]. This process usually results in a reliable ranking of ideas and users who generated them. However, special means are needed in order to develop a deeper understanding of users's behavior and to perform complex dynamic and statistic analyses. Traditional methods of clustering, community detection and text mining should be adapted or even fully redesigned. Moreover, these methods require ingenuity for their effective and efficient use that will allow to find non-trivial results. We try to bridge this gap and propose new methodology that can be helpful for these crowdsourcing platforms.

This paper extends on our previously published paper [4] and is devoted to the modeling of a recommender system for crowdsourcing platforms. In the previously published paper we have already described the general methodology, however, this paper provides further details as well as through experiments. The paper has the following structure: Section 2 provides more details about Witology platform, Section 3 briefly describes FCA-based data representations and methods, Section 4 discusses the selected results of the experiments, Section 5 concludes the paper.

2 Witology platform

For the better understanding of the problem, we provide more details about the Witology crowdsourcing project. The crowdsourcing process at this platform features eight main stages: "Solution's generation", "Selection of similar solutions", "Generation of counter-solutions", "Total voting", "Solution's improvements", "Solution's stock", "Final improvements" and finally "Solution's review".

The first stage "Solution's generation" is performed individually by each user. A key difference between the traditional brainstorming and the "Solution's generation" stage is that the participant are not aware of the ideas of other participants. The main similarity is the absence of criticism which was postponed till the later stages. In the "Selection of similar solutions" phase participants are

[12] http://sberbank21.ru/
[13] http://www.asi.ru/en/

selecting similar ideas (solutions) and their aggregated opinions are transformed into clusters of similar ideas.

Counter-solutions generation includes criticism (pros and cons) and evaluation of the proposed ideas by means of communication between an author and experts. During this stage author of the idea can invite other experts to his team taking into account their contribution to the discussion and criticism. "Total voting" is performed by evaluating each proposed idea by all the users where votes indicate users' attitude as well as quality of the proposed solution (marks are integers between -3 and 3). Two stages, i.e. "Solution's improvements" and "Final improvements", involve active collaboration of experts and authors who work to improve their solutions together.

"Solution's stock" is one of the most interesting game stages of the project. At this stage all the participants who were able to accumulate positive reputation at the previous stages receive money in internal currency "wito" and can take part in stock trade.

The way crowdsourcing platforms work is very different from the principles underlying online-shops or specialized music/films recommender sites. Crowdsourcing projects consist of several stages where the results of each stage greatly depend on the results of a previous one. This is the reason why existing recommender models should be adapted considerably. We have developed new methods for making recommendations based on well-known mathematical approaches and propose methods for idea recommendation (for voting), like-minded persons recommendation (for collaborating) and antagonists recommendation (for contridea generation stage). To the best of our knowledge, it is the first time the last recommendation type is proposed.

3 FCA-based models for crowdsourcing data

At the initial stage of analysis of the data acquired from the collaborative platform, two data types were identified: data without keywords (links, evaluations, user actions) and data with keywords (all user-generated content).

For the analysis of the 1st type of data (without keywords) we applied Social Network Analysis (SNA) methods, clustering (biclustering and triclustering [5,6], spectral clustering), Formal Concept Analysis (FCA) [7] (concept lattices, implications, association rules) and its extensions for multimodal data, triadic, for instance [8]; recommender systems [9,10] and statistical methods of data analysis [11] (the analysis of distributions and average values).

Basic definitions of FCA and OA-biclustering can be found in [4].

It is easy to show that all key crowdsourcing platform data can be described in FCA terms by means of formal contexts (single-valued, multi-valued or triadic).

1. The data below is described by a single-valued formal context $\mathbb{K} = (G, M, J)$.

Let $\mathbb{K}_P = (U, I, P)$ be a formal context, where U is a set of users, I is a set of ideas, and $P \subseteq U \times I$ shows which user proposed which idea. Two other contexts, $\mathbb{K}_C = (U, I, C)$ and $\mathbb{K}_E = (U, I, E)$, describe binary relations of idea commenting and idea evaluation respectively.

The user-to-user relationships can also be represented by means of a single-valued formal context $\mathbb{K} = (U, U, J \subseteq U \times U)$, where $u_1 J u_2$ can designate, for example, that user u_1 commented some idea proposed by u_2. Relationships between content items can be modelled in the same way, e.g. $\mathbb{K} = (T, T, J \subseteq T \times T)$, where $t_1 J t_2$ shows that t_1 and t_2 occurred together is some text (idea or comment).

2. A multi-valued context $\mathbb{K}^W = (G, M, W, J)$ can be useful for representing data with numeric attributes.

Let $\mathbb{K}^F = (U, K, F, J)$ be a multi-valued context, where U is a set of users, K is a set of keywords, F is a set of keyword frequency values, $J \subseteq U \times K \times F$ shows how many times a particular user u applied a keyword k in an idea description or while discussing some ideas. The context $\mathbb{K}^F$ can be reduced to a plain context by means of (plain) scaling.

The commenting and evaluation relations can be described through multi-valued contexts in case we count each comment or evaluation for a certain topic. E.g, the multi-valued context $\mathbb{K}^V = (U, I, V = \{-3, -2, -1, 0, 1, 2, 3\}, J)$ describes which mark a particular user u assigns to an idea i, where V contains values of possible marks; it can be written as $u(i) = w$, where $v \in V$.

For more information about triadic models see [4].

3.1 FCA-based recommender model

Two kinds of recommendations seem to be potentially useful for crowdsourcing. The first one is a recommendation of like-minded people to a particular user, and the second one is the ability to find antagonists, users which discussed the same topics as a target one, but with opposite marks.

1. **Recommendations of like-minded persons and interesting ideas.**

Let $\mathbb{K}_P = (U, I, P)$ be a context which describes idea proposals. Consider a target user $u_0 \in U$, then every formal concept $(A, B) \in \mathfrak{B}_P(U, I, P)$ containing u_0 in its extent provides potentially interesting ideas to the target user in its intent and prospective like-minded persons in $A \setminus \{u_0\}$.

Consider the set $\mathfrak{R}(u_0) = \{(A, B) | (A, B) \in \mathfrak{B}_P(U, I, P)\ and\ u_0 \in A\}$ of all concepts containing a target user u_0. Then the score of each idea or user to recommend to u_0 can be calculated as follows $score(i, u_0) = \frac{|\{u | u \in A,\ (A,B) \in \mathfrak{R}(u_0)\ and\ i \in u'\}|}{|\{u | u \in A\ and\ (A,B) \in \mathfrak{R}(u_0)\}|}$ or $score(u, u_0) = \frac{|\{A | u \in A\ and\ (A,B) \in \mathfrak{R}(u_0)\}|}{|\mathfrak{R}(u_0)|}$ respectively. As a result we have a set of ranked recommendations $R(u_0) = \{(i, score(i)) | i \in B\ and\ (A, B) \in \mathfrak{R}\}$. One can select the topmost N of recommendations from R ordered by their score.

2. **Recommendations of antagonists.**

Consider two evaluation contexts: the multi-valued context $\mathbb{K}^W = (U, I, W = \{-3, -2, -1, 0, 1, 2, 3\}, J)$ and binary context $\mathbb{K}_E = (U, I, E)$. Then consider (X, Y) from $\mathfrak{R}(u_0) = \{(A, B) | (A, B) \in \mathfrak{B}_P(U, I, P)\ and\ u_0 \in A\}$. Set X contains people that evaluated the same set of topics Y, but we cannot say that all of them are like-minded people w.r.t relation E. However, we can introduce a distance

measure, which shows for every pair of users from X how distant they are in marks of ideas evaluation:

$$d_{(X,Y)}(u_1, u_2) = \sum_{\substack{u_1,u_2 \in X \\ i \in Y}} |i(u_1) - i(u_2)|. \quad (1)$$

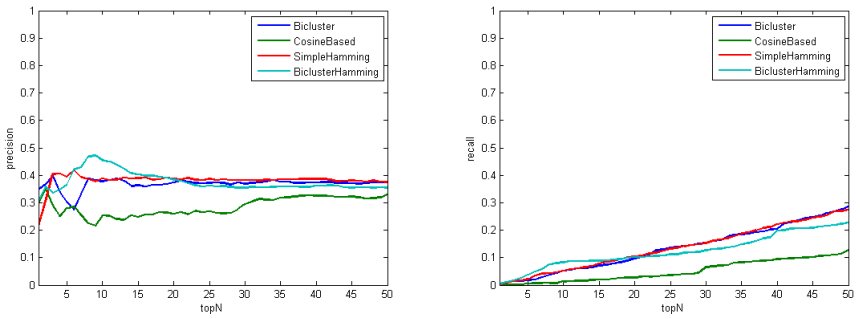
As a result we again have a set of ranked recommendations $R_{(X,Y)}(u_0) = \{(u, d(i))|u \in U \text{ and } (A, B) \in \mathfrak{R}\}$. The topmost pairs from $R_d(u_0)$ with the highest distance contain antagonists, that is persons with the opposite views on most of the topics which u_0 evaluated. To aggregate $R_{(X,Y)}(u_0)$ for different (X, Y) from $\mathfrak{R}(u_0)$ into a final ranking we can calculate

$$d_{(u_0,u)} = \max\{d_{(X,Y)}(u_0, u)|(X, Y) \in \mathfrak{R}(u_0) \text{ and } u_0, u \in X\}. \quad (2)$$

The proposed models were tuned and validated, and they also had several variations such as using biclusters instead of formal concepts and other ways of final distance calculation.

4 Results of the experiments

We proposed and implemented several methods for antagonists recommendation task: bicluster-based, cosine (or correlation) based, simple and bicluster-based Hamming methods. In Figure 1 we can see that for small Top-N output the best result is achieved by biclustering with Hamming metric, but for large Top-N output both methods, biclustering with correlation and simple Hamming distance, show almost equally good results. Taking into account that our goal was to reduce user's time spending on crowdsourcing tasks, we need small Top-N, and the best choice is biclustering with Hamming distance (almost 0.5 precision).



(a) (b)

Fig. 1. Antagonists recommendation precision (a) and recall (b)

5 Conclusion

The paper presented a new methodology that can be applied to the data acquired from crowdsourcing platforms. The results of our experiments suggest that the developed methods are useful for making recommendations in the Witology crowdsourcing system and able to support user's activity on the platform.

Acknowledgments. The main part of this work was performed by the project and educational group "Algorithms of Data Mining for Innovative Projects Internet Forum". Further work was supported by the Basic Research Program at the National Research University Higher School of Economics in 2012-2014 and performed in the Laboratory of Intelligent Systems and Structural Analysis. First author was also supported by Russian Foundation for Basic Research (grant #13-07-00504)

References

1. Roth, C., Cointet, J.P.: Social and semantic coevolution in knowledge networks. Social Networks **32** (2010) 16–29
2. Yavorsky, R.: Research Challenges of Dynamic Socio-Semantic Networks. In Ignatov, D., Poelmans, J., Kuznetsov, S., eds.: CEUR Workshop proceedings Vol-757, CDUD'11 - Concept Discovery in Unstructured Data. (2011) 119–122
3. Howe, J.: The rise of crowdsourcing. Wired (2006)
4. Ignatov, D.I., Kaminskaya, A.Y., Bezzubtseva, A.A., Konstantinov, A.V., Poelmans, J.: Fca-based models and a prototype data analysis system for crowdsourcing platforms. In Pfeiffer, H.D., Ignatov, D.I., Poelmans, J., Gadiraju, N., eds.: ICCS. Volume 7735 of Lecture Notes in Computer Science., Springer (2013) 173–192
5. Barkow, S., Bleuler, S., Prelic, A., Zimmermann, P., Zitzler, E.: Bicat: a biclustering analysis toolbox. Bioinformatics **22**(10) (2006) 1282–1283
6. Ignatov, D.I., Kuznetsov, S.O., Poelmans, J., Zhukov, L.E.: Can triconcepts become triclusters? International Journal of General Systems **42**(6) (2013) 572–593
7. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. 1st edn. Springer-Verlag New York, Inc., Secaucus, NJ, USA (1999)
8. Jäschke, R., Hotho, A., Schmitz, C., Ganter, B., Stumme, G.: TRIAS–An Algorithm for Mining Iceberg Tri-Lattices. In: Proceedings of the Sixth International Conference on Data Mining. ICDM '06, Washington, DC, USA, IEEE Computer Society (2006) 907–911
9. Ignatov, D.I., Kuznetsov, S.O.: Concept-based Recommendations for Internet Advertisement. In Belohlavek, R., Kuznetsov, S.O., eds.: Proc. CLA 2008. Volume Vol. 433 of CEUR WS., Palacky University, Olomouc, 2008 (2008) 157–166
10. Ignatov, D.I., Poelmans, J., Dedene, G., Viaene, S.: A New Cross-Validation Technique to Evaluate Quality of Recommender Systems. In Kundu, M.K., Mitra, S., Mazumdar, D., Pal, S.K., eds.: PerMIn. Volume 7143 of LNCS., Springer (2012) 195–202
11. Clauset, A., Shalizi, C.R., Newman, M.E.J.: Power-law distributions in empirical data. SIAM Rev. **51**(4) (November 2009) 661–703