2008 Third International Workshop on Semantic Media Adaptation and Personalization

Modeling User Networks in Recommender Systems

Dimitrios Vogiatzis
Institute of Informatics and Telecommunication
NCSR "Demokritos" Athens, Greece
dimitrv@iit.demokritos.gr

Nicolas Tsapatsoulis
Department of Communication and Internet Studies
Cyprus University of Technology, Limassol, Cyprus
nicolas.tsapatsoulis@cut.ac.cy

Abstract

Recommender systems, in the collaborative filtering variation, are popular tools used to drive users out of information clutter, by letting them select "interesting" items based on the preferences of similarly minded users. In such a system as more users come in to evaluate items (be they information pieces, products or otherwise), a network of users starts to be formed. In this paper we are interested in the dynamics of such a network, in particular we investigate if there is a hidden law that captures the essence of such networks irrespective of their size. The discovery of such a law would allow, among other usages, generation of synthetic data sets, realistic enough to be used for simulation purposes. Furthermore, it would be useful for information-seeking activities such as locating known experts or influential users on a particular subject. Similar work in related fields suggested the existence of power-laws, which seem to be ubiquitous. However, in our work we did not detect the presence of such a law, instead we discovered an exponential relationship between the nodes of a graph representing users, and edges representing similarity between users. In particular the logarithm of the degree of node is linearly related to the ranking of the node in a decreasing order. The above conclusion is justified by extended experiments on two versions of the *movie lens* data set (one comprised 100,000 user evaluations, while the other comprised 1,000,0000 evaluations).

1 Introduction

As e-commerce proliferates, recommender systems seem also to expand. Recommender systems aim to "promote" items of possible interest to a user, thus driving him out of the information clutter. An important component of a recommender system is user modeling, which allows to compute common preferences of the users, and thus to *link* people that have got something in common. Once enough users have joined such a system, a network of linked people starts to be formed. There are many strands of research in recommender systems including collaborative filtering and content base systems (see [2] for an overview). In collaborative filtering, a community of users recommends items of possible interest to the active user. The active user receives recommendations, based on his/her similarity to the members of this community. Similarity, between two users can be defined in terms of the similar evaluations they have provided on the same items in the past. Thus, a new item, unknown to the active user, but been positively evaluated by the community can be suggested to the active user.

Recommendations, could take advantage of the tremendous potential of such a social network that is self formed and self organised. Such recommendations would be a step forward than the aforementioned recommendations strategies. A survey in recommender systems and social networking can be found in [7] while in [3] there is a seminal talk which raises issues about data collection, analysis

978-0-7695-3444-2/08 $25.00 © 2008 IEEE
DOI 10.1109/SMAP.2008.35

and modeling of the web graph.

In the current work we aim to model the network that is formed by users that rate movies they have seen. Such users, based on their similarities, can form a social network. Furthermore, such a network is highly dynamic: new users join the network and existing users rate new movies. Therefore the network can be modeled as a graph that evolves over time. The issues we address is whether there is a pattern in the network and whether the network reaches a stable form irrespective of its size.

There is significant research in the search of power laws in internet topology, and in the influence of users in recommender systems. However, the network *per se* of users in a recommender system has not been explored yet.

In [5] there is study of the internet topology between late 1997 and late 1998 which detected the presence of power laws. In particular such a relationship holds between the outdegree of nodes and their rank, where the *outdegree* of a node is defined as the number of edges incident to that node and the *rank* refers to its index in the order of decreasing outdegree. An extension of this study can be found in [10].

Concerning recommender systems, the authors in [9] studied the influence of individual users in collaborative recommendation systems. The formation of a recommendation depends heavily on some users' past evaluations, hence the most influential users drive the recommendation. A power-law relationship was discovered between the influential users and their ranking in a decreasing influence order.

A similar study was conducted in [4] to discover the *network value* of customers, which is defined as the marketing value of a customer in terms of other customers that maybe influenced by him. In a collaborative filtering recommender system the network value of individual users and their ranking in a decreasing network value is also governed by a power-law.

The study of dynamics of collaborative tagging also reveals a power law [6]. There are sites such as "del.icio.us" that offer visitors the service of characterising a site through tags. It was discovered that the collection of all tags and their frequencies ordered by rank frequency for a given resource after some time tends to be stable and to follow a power law distribution.

The rest of paper is organised as follows, in Sect. 2 we formalise the problem we are addressing, then in Sect. 3 we refer to the experiments we have performed, and in Sect. 4 we expose the results followed by a relevant discussion. Finally, in Sect. 5 we draw conclusions and directions for future work.

2 Modeling Methodology

The recommendation problem can be formulated as follows: Let C be the set of all users and let I be the set of all possible items that the users can recommended, such as books, movies, or restaurants. Let also u be a utility function that measures the usefulness (as may expressed by user ratings) of item i to user c, i.e., $u : C \times I \rightarrow R$. The usefulness of all items to all users can be expressed as a matrix U with rows corresponding to users and columns corresponding to items. An entry $u(c, i)$ of this matrix may have either positive value indicating the usefulness (rating) of item i to user c or a zero value indicating that the usefulness $u(c, i)$ has not been evaluated. The recommendation problem can be seen as the estimation of zero values of matrix U from the non-zero ones.

Recommendation in the collaborative filtering approach requires some similarity $r(a, b)$ between users a and b to be computed based on the items that both of them evaluated with respect to their usefulness. We can define a user similarity matrix R with entries $r(a, b)$, $a, b \in S$. Various approaches have been used to compute the similarity between users in collaborative recommender systems. In most of these approaches, the similarity between two users is based on their ratings of items that both users have evaluated. The two most popular approaches are correlation and cosine-based. Both of these methods produce values $r(a, b) \in [0\ 1]$. Zero values of matrix R may correspond to either zero similarity, or, to users with no commonly evaluated items. Therefore, users with many evaluations are likely have many similarities among other users. These users are usually called "influential users". The influence of a user can be computed by taking the sum across the corresponding row or column of matrix R. The higher this sum is the more influential the user is.

In our approach, we are interested in representing the concept of similar users as a graph, where nodes will represent users. If two users are similar enough, an edge will connect their respective nodes. Thus the number of edges stemming out of a user's node (i.e. its degree) will de-

note the number of its similar users. The edges are bi-directional, denoting that if user a is similar to user b then the reverse also holds.

As usual, similarity between two users is based on the items both have evaluated. Let a and b be two users and k is the number they both have evaluated (either positively with 1 or negatively with 0). User similarity is based on hamming distance H between their representative vectors, normalised by their length. Let $d(a_j), d(b_j) \quad \forall i, j$ the evaluations of user a, b on item j,

$$a = [d(a_1), d(a_2), \dots d(a_k)] \tag{1}$$

$$b = [d(b_1), d(b_2), \dots d(b_k)] \tag{2}$$

their similarity s is defined as,

$$s = 1 - H(a, b)/\|a\| \tag{3}$$

where $s \in [0, 1]$, with higher values denoting closer similarity.

The degree of a node a is defined as,

$$degree_a = length[b \in M, \quad \forall b : s(a, b) > t] \tag{4}$$

where b denotes a node, M is the set of all the nodes and t is a threshold value, with $t \in (0, 1)$.

We are interesting in examining whether a power-law holds between the degree of the node and its ranking, in decreasing degree. The existence of a power law would state, that very few users share the same preferences with many others, whereas the vast majority share the same preferences with very few. Furthermore, we were interested to see if this phenomenon holds over time. That is when new users join the system to offer recommendations, or existing users offer more recommendations, then what is the effect of that? Also, the discovery of a power law would enable people to generate synthetic data sets for simulation purposes, that are realistic enough.

Moreover, a new strand of research in collaborative base recommendation systems has introduced the concept of trust, which states that some user when recommending are to be trusted more than others. Trusted users have some important characteristics, for instance, they have evaluated many items and consequently they are "similar" to many other users. Thus a study similar to the above might suggest a rough estimate of the trusted users that are expected to be found, or that indeed worth to be found. Also in [1] there is a proposal for search in power-law graphs, which increases sublinearly as the graph grows by concentrating on highly connected nodes.

A power law distribution satisfies the following [11]:

$$p(X \geq x) = kx^{-a} \tag{5}$$

where $a > 0$. The probability density function is derived as

$$f(x) = ak^a x^{-a-1} \tag{6}$$

Since we interested in the case where X is discrete then

$$p(X = x) = kx^{-a} \tag{7}$$

thus $log[p(X = x)] = (-a)log(x) + log(k)$.

The characteristic property of a power law distribution is the heavy tail, which tells that the distribution decreases slowly, in contrast the normal distribution decreases exponentially fast. Also, power-law distribution are scale free distributions, in the sense that if x is multiplied by a factor the distribution still holds.

3 Experimental Setting

We have used the movie lens data set [1], a widely used data set. It is comprised of users (some demographic data are also provided for each user), which have seen and evaluated movies. In addition for each movie, there is information about the categories it belongs to (e.g. adventure, police etc.) as well as various other pieces of information. In this data set there are 100,000 evaluations on 1 to 5 (5 denoting the highest evaluation) scale on 1682 movies by 943 users (henceforth we will call it *small*). The data were recorded from late September of 1997 to late April of 1998.

We have split the data into four time periods, to examine the time evolution of the network of similar users. The first period starts at the beginning, and ends at the end of first quarter of the data. The second period starts at the beginning at ends at the second quarter of the data, and so on until the fourth period.

[1] GroupLens research lab, Department of Computer Science and Engineering, University of Minnesota (http://www.grouplens.org/taxonomy/term/14)

We repeated the same series of experiments with a larger version of the aforementioned dataset, provided by the same source. It contained 1,000,000 evaluations on 1 to 5 scale on 3952 movies from 6040 users. The data were recorded from late April 2000 till the end of Feb in 2003 (henceforth we will call it *big*). Both data sets contained users that have evaluated more than 20 movies.

We have preprocessed both data sets, by condensing the ratings of 1,2 to 0 and 3,4,5 to 1; denoting approval or disapproval of a movie respectively. Consequently, similarity between two users being based on movies the have seen and rated , is measured as state in a previous section. Also, we have ordered the data according to the time they where obtained.

4 Results and Discussion

In figures 1, and 2 the results from the small and large data sets appear. The y axis represents, the degree of a node, and the x axis represents the ranking of the node in an order of decreasing degrees. The y-axis is on log_{10} scale. Also, the nodes that have zero degree have been removed. The line that appears between the dots, is produced through linear regression [8].

A relation of the form

$$log_{10}(y) = a * n + b \tag{8}$$

is revealed, where $n \in N$ and a, b are the line parameters. Thus an exponential relation between the degree of a node y and its ranking n. The regression results appear in Table 1 for the four time frames for the two data sets. The above denote that many nodes (representing users) have a low degree, where as a few nodes have high degree. (An edge connecting two nodes, denoted that these nodes are close enough in terms of past movie ratings to be considered as users). Thus we did not find a power-law relationship, instead the expected heavy tail drops sharply. In Tables 2 and 3 there is information about the mean degree of the nodes, the number of nodes, and the mean square error of the linear regression approximation to the data points, for both data sets and for the four periods. The value of the threshold t mentioned in Sect. 2 has been set to 0.06. It was experimentally determined in the range of $[0.09, 0.04]$ the exponential relation of Eq. 8 holds. Higher values produce a very sparsely connected graph and lower values a densely connected graph effectively destroying the concise form of equation 8.

Table 1: Linear Regression equations

	Small	Large
period-1	-0.0597*n+1.21	-0.0220*n+1.88
period-2	-0.0237*n+1.80	-0.0096*n+2.30
period-3	-0.0146*n+2.01	-0.0067*n+2.35
period-4	-0.0113*n+2.24	-0.0037*n+2.61

Table 2: Small data set, Network characteristics

	mean degree	number of users per period	MSE
period-1	5.14	280	0.0080
period-2	13.72	491	0.0037
period-3	20.81	708	0.0061
period-4	31.86	943	0.0140

5 Conclusions and Future work

Social computing applies computational techniques to the study of social interactions. In the current work we were interested in modeling the network of similar users in a setting that is typically used in collaborative filtering recommender system. A prominent role in our study had the *movie lens* which contained movie evaluations by users in two different data sets.

We represented the network of similar users as a graph,

Table 3: Big data set, Network characteristics

	mean outdegree	number of users per period	MSE
period-1	15.93	1772	0.0041
period-2	36.12	3255	0.0091
period-3	38.58	5140	0.0036
period-4	64.26	6040	0.0028

where nodes denote users and an edge connects two users if they are similar enough. It was discovered a linear relationship between the logarithm of the degree of a node y and its ranking in a decreasing degree order n, that is $log_{10}(y) = a * n + c$, where a and c are constants and $n \in N$. Contrary to what we might expect from relevant literature a power-law was not discovered.

Also, the study aimed to observe the evolution of the user's network in time, in particular what are is the form it assumes as it evolves in time. For that reason we studied the networks' form in four time instances. From the experimental results it is derived that the law described by Eq. 8 is established, even at the earlier stages. However, the slope of the line seems to vary as new users or old users produce more evaluations.

In the future, we intend to investigate whether the discovered law reaches a stable form, that is whether the a and c parameters reach a certain value or they vary in a chaotic way as more users are added. Also, the overall study, was based on a similarity measure, which depends on a threshold. Similarity between two users above the threshold induces a edge between them. The higher the threshold the more sparse the graph becomes. Further study on the validity of the threshold is required, especially because a large number of nodes is left with no edges, that is with zero degree. Moreover, we did not study whether there are disjoint components in the graph of users and whether a different set parameters for the discovered law would hold for each component. Finally, the failute to discover a power law came as a surprise, since in physics, biology, geology, and of course in computer and social networks power-laws seem to be ubiquitous (see also the literature review in the introduction). Thus we intend to further investigate, whether network size, network connectivity and other parameters such as the definition of similar users may affect the modeling.

References

[1] L. A. Adamic, R. M. Likose, A. R. Puniyani, and B. A. Huberman. Search in power-law networks. *Physical Review E*, 2001.

[2] G. Adomavicius and A. Tuzhilin. Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6), 2005.

[3] A. Broder. Keynote address - exploring, modeling, and using the web graph. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003.

[4] P. Domingos and M. Richardson. Mining the network value of customers. In *KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 57–66, New York, NY, USA, 2001.

[5] M. Faloutsos, P. Faloutsos, and C. Faloutsos. On Power-Law Relationships of the Internet Topology. In *Proceedings of the conference on Applications, technologies, architectures, and protocols for computer communication*, 1999.

[6] H. Halpin, V. Robu, and H. Shepherd. The complex Dynamics of Collaborative Tagging. In *Proceedings of the 16th international conference on World Wide Web*, 2007.

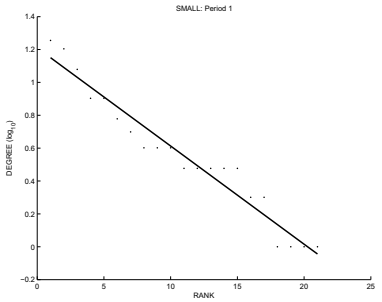
[7] S. Perugini, M. Gonçalves, and E. Fox. A Recommender Systems Research: A connection-Centric Survey. *Journal of Intelligent Information Systems*, 2004.

[8] W. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge University Press, third edition, 2007.

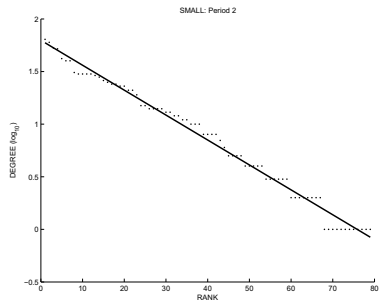
[9] M. Rashid, G. Karypis, and J. Riedl. Influence in Ratings-Based Recommender Systems: An Algorithm-Independent Approach. In *SIAM Proceedings of International Conference on Data Mining*, 2005.

[10] G. Siganos, M. Faloutsos, P. Faloutsos, and C. Faloutsos. Power-Laws and the AS-level Internet topology. *ACM/IEEE Transactions on Networking*, pages 514–524, 2003.

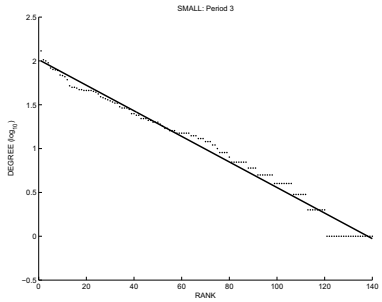
[11] W. Willinger, V. Paxson, and M. Taqqu. Self-similarity and heavy tails: Structural modeling of network traffic. In J. Adler, R. E. Feldman, and M. S. Taqqu, editors, *Statistical Techniques and Applications*. Birkhauser, 1998.



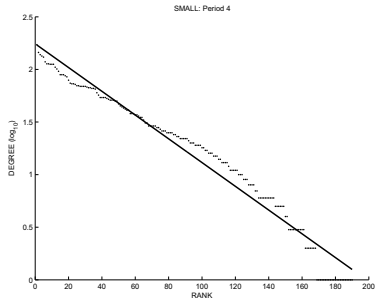
(a) 20 Sep 1997 - 13 Nov 1997



(b) 20 Sep 1997 - 22 Dec 1997

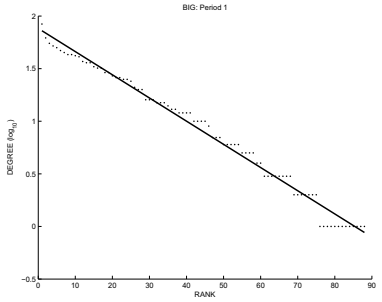


(c) 20 Sep 1997 - 23 Feb 1998

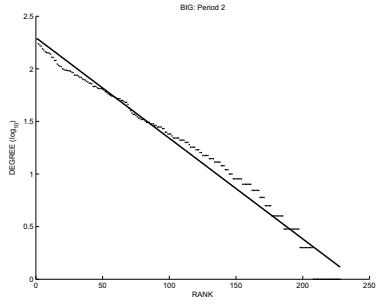


(d) 20 Sep 1997 - 22 Apr 1998

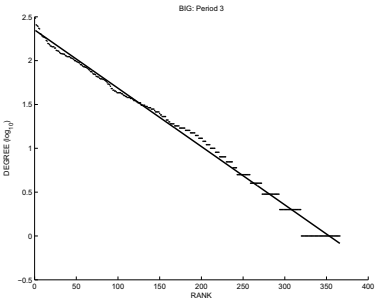
Figure 1: Small data set



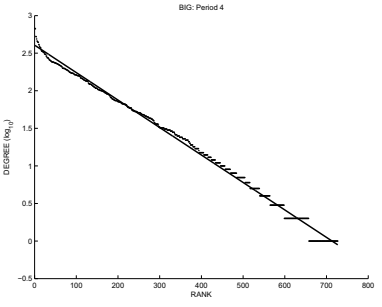
(a) Tue, 25 Apr 2000 - 03 Aug 2000



(b) 25 Apr 2000 - 31 Oct 2000



(c) 25 Apr 2000 - 26 Nov 2000



(d) 25 Apr 2000 -28 Feb 2003

Figure 2: Big data set