A Social Network Study of Evliyâ Çelebi's *The Book of Travels-Seyahatnâme*

Ceyhun Karbeyaz[*]
Bilkent University
Department of Computer Engineering
Ankara 06800, Turkey
karbeyaz@bilkent.edu.tr

Ethem F. Can[†]
Bilkent University
Department of Computer Engineering
Ankara 06800, Turkey
ethemfatih.can@gmail.com

Fazli Can
Bilkent University
Department of Computer Engineering
Ankara 06800, Turkey
canf@cs.bilkent.edu.tr

Mehmet Kalpakli
Bilkent University
Department of History
Ankara 06800, Turkey
kalpakli@bilkent.edu.tr

ABSTRACT

Evliyâ Çelebi, an Ottoman writer, scholar and world traveler, visited most of the territories and also some of the neighboring countries of the Ottoman Empire in the 17^{th} century. He took notes about his trips and wrote a 10-volume book called *Seyahatnâme (The Book of Travels)*. In this paper, we present two methods for constructing social networks by using textual data and apply it to *Seyahatnâme-Bitlis Section* from book IV and check if the constructed networks hold social network properties. The first social network construction method is based on proximity of co-occurence of names. The second method is based on 2-pair associations obtained by association rule mining by using sliding text blocks as transactions. The social networks obtained by these two methods are validated using a Monte Carlo approach by comparing them with the social network created by a scholar-historian.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: [Data mining]

General Terms

Experimentation, Verification

[*]Current Address: Department of Electrical and Computer Engineering, Northeastern University, 02115 USA

[†]Current Address: Department of Computer Science, University of Massachusetts Amherst, 01003 USA

ENRICH 2013 Workshop, August 1, 2013, Dublin, Ireland.

Keywords

Social networks, data mining, historical documents

1. INTRODUCTION

Evliyâ Çelebi; a 17^{th} century Ottoman writer, scholar, and world traveler (born on 1611, died circa 1682); visited most of the territories and also some of the neighboring countries in Africa, Asia and Europe of the Ottoman Empire over a period of 40 years. His work *Seyahatnâme (The Book of Travels)* is known by its distinguished style and detailed descriptions of people and places that he visited during his long journeys [4]. A preliminary version of this study can be seen in [10].

Seyahatnâme is a long masterpiece that is composed of ten books each containing information about different locations and several prominent people of its time. This nature of the work is appealing for social network studies which aim to identify relationships among people. He described the people and the incidents he came across in Bitlis in book IV and V. In our study we aim to automatically identify relationships among the people who appear in the text. These identified relationships between the important historical characters would open new research avenues [12]. For this purpose we use the text in transcribed form [9] and perform several experiments.

The contributions of this study can be summarized as follows. We present two different methods for constructing social networks from textual data and apply them to *Seyahatnâme-Bitlis Section* to obtain a social network that represent relationships among people. The first social network construction method we present is based on proximity of co-occurence of names. The second method is based on 2-pair associations obtained by association rule mining [1]. We use the social network created by a human expert as the ground truth and assess the effectiveness of the methods by comparing the generated network structure with that of the ground truth. We use a Monte Carlo approach for validating the automatically constructed social networks and show that the social network structures obtained by our methods are significantly different from random. We also analyze the manually and automatically generated networks to see if they contain the social network properties.

2. RELATED WORK

One of the methods we present in this study is based on association rules, which are the derived relations between the items of a dataset. The notion was first introduced by Agrawal et al. [1] and later the notion of assocation rules are used in many studies. In this method, let I = $\{i_1, i_2, ..., i_n\}$ be a set of items with size n and T = $\{t_1, t_2 ,..., t_m\}$ is set of transactions (in market data analysis a transaction involves the group of items purchased together) with size m. Then an association rule is shown as $X \Rightarrow Y$ where $X, Y \subseteq I$ and $X \cap Y = \oslash$. The aim of the association rules is to find all pairs of items within all T that have support S and confidence C values greater than user defined amounts. S is a threshold to extract the frequent itemsets from transactions whereas C is the ratio of support for items X and Y occuring together to support of item X. The best-known algorithm to mine association rules is Apriori [2].

Brin et al. [3] use association rule mining to find frequent itemsets over market data. They show that their algorithm provides better performance with respect to the Apriori algorithm while finding large itemsets. Raeder and Chawla [13] applied association rules over construction and analysis of a social network from market basket data. Their main goal was to search for meaningful relationships over the formed social network. They conclude that the highly rated product communities have a significant relationship with a clear purpose in the social network.

3. METHODS: ProxiBM and RuleBM

In this paper, we present two social network construction methods. These are the text proximity-based method (ProxiBM) and the association rule-based method (RuleBM). Both methods are based on co-occurrence of names in close proximity within a text block.

For determing text blocks with a meaningful cohesive context we use two approaches. In ProxiBM we use the paragraph information provided in the transcribed text. We manually identified and tagged 164 paragraphs. Each paragraph is used as a block. However, in the original work of Evliyâ Çelebi there are no explicit paragraphs. Furthermore, manual paragraph tagging is difficult. This provides a motivation for automatic identification of blocks, hence in RuleBM we employ a sliding text window approach and use a certain number of consecutive words as a block. The obtained blocks partially overlap. Since blocks are overlapping and obtained from consecutive words we intuitively expect that they will at least have a (partially) cohesive context. The sliding blocking approach is explained in Figure 1.

In ProxiBM, edges for the undirected graph of social network are derived by creating a link between every character that appear in the same paragraph within a close word proximity (using a threshold). The proximity threshold between any two names is varied between 5 words to 500 words in steps. This approach is inspired by the use of term closeness as an indicator of document relevance [6] and by an earlier study that uses a similar approach in social network analysis [14].

In RuleBM we use the sliding text window for blocking and treat each block as a transaction. The names that appear in a block correspond to shopping items; in this way, we are able to extract association rules [1]. By using the Apriori algorithm[2] we derive the 2-item pairs from the transactions by using a support threshold value. (Note that in RuleBM association rules that involve more than two character names are too few to use.) In RuleBM we use the 2-pair association rules as relational edges of the social network. We employ different support threshold values and repeat the blocking operation for different block and sliding window sizes in order to find the best performing parameters.

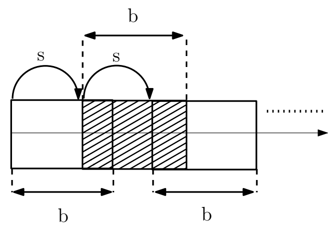

Figure 1: Sliding window-based blocking: l: total text length, b: block size ($0 < b \leq l$), s: step size, nb: number blocks, $nb = 1 + \lceil \frac{(l-b)}{s} \rceil$ for $0 < s \leq b$, $nb = 1 + \lfloor \frac{(l-b)}{s} \rfloor$ for $s > b$.

4. MEASURING EFFECTIVENESS

For measuring the effectiveness of the methods in predicting the correct links we use the social network created by Prof. Kalpaklı, a scholar-historian. The agreement between automatically constructed social networks and the manually constructed (actual) social network is measured by *precision*, *recall*, and the *F-measure* [11] (their formulas are given below). In *precision* we calculate what fraction of of automatically generated links are correct. In *recall* we calculate what fraction of actual links is identified. *F-measure* is a harmonic mean of these two measures [11]. They all assume a value between 0 (worst case: no match at all) and 1 (best case: perfect match). While constructing the links only the people of that time are considered, i.e., names of people who were not alive, prophet names, different names of god etc. are excluded.

$$Precision = \frac{No.\ of\ matching\ edges}{No.\ of\ edges\ obtained\ by\ method} \quad (1)$$

$$Recall = \frac{No.\ of\ matching\ edges}{No.\ of\ edges\ of\ manually\ constructed\ network} \quad (2)$$

$$F = \frac{2 * Precision * Recall}{Precision + Recall} \quad (3)$$

5. EXPERIMENTAL RESULTS

5.1 Generation and Validation of Social Networks

In the experiments we first measure the accuracy of automatically generated social network by measuring their similarity with the ground truth data using *precision*, *recall*, and *F-measure*. In the experiments, both methods are tested in various conditions in order to find their best matching (most similar) configuration to the ground truth. For ProxiBM we use various proximity threshold values. For RuleBM we report the results for block sizes 500 and 1,000 words (with block size smaller than 500 and larger than 1,000 we obtain low effectiveness values and we do not report them here). Table 1 shows the *precision*, *recall* and *F-measure* results of ProxiBM for different proximity threshold values. The best configuration for this method with and *F-measure* value of 0.59 is observed when proximity threshold is 25 words.

Table 1: Performance results of ProxiBM over paragraphs for different proximity threshold (Θ) values in terms of no. of words.

Measure	$\Theta = 5$	$\Theta = 10$	$\Theta = 25$	$\Theta = 50$	$\Theta = 100$	$\Theta = 250$	$\Theta = 500$
Precision	0.47	0.52	0.54	0.49	0.48	0.46	0.45
Recall	0.16	0.39	0.66	0.70	0.70	0.71	0.71
F-measure	0.24	0.44	0.59	0.58	0.57	0.56	0.56

A similar experiment is done for RuleBM that is based on assocation rule mining. For this purpose, exactly the same blocking operation is done by considering sliding window sizes. And again the experiment is performed for block sizes 500 words and 1,000 words. Association rules are derived from these blocks for different support thresholds ranging from 5% to 20%. The best configuration for this method with and *F-measure* value of 0.29 is observed when blocksize: 500 words, stepsize: 300 words and support threshold: 5%.

Another set of experiments are conducted that aim to understand if the automatically generated social networks are significantly different from random. For this purpose Monte Carlo experiments are performed [8]. Monte Carlo experiments define a reference population and provide a baseline distribution [8, pp. 161-162]. Note that all randomly created social networks that are used in our tests have the same network properties with the automatically generated network that is being evaluated (e.g., the same number of nodes and the same average degree distribution within the nodes where degree of a node is the number of incoming edges to that node). In order to achieve this the Erdős-Renyi random network generation algorithm is used [5]. In all Monte Carlo experiments, we generate a random version of the social network which is being evaluated 1,000 times and measure the average *F-measure* values. The reported *F-measure* values are obtained from the averages obtained for 1,000 *precision* and *recall* observations. The comparison of average *F-measure* results for the proposed methods versus Monte Carlo average *F-measure* results for different experimental conditions are provided in Figure 2 and Figure 3 for ProxiBM and RuleBM, respectively. The plots show that both methods with proper parameters generate networks which are significantly different from random.

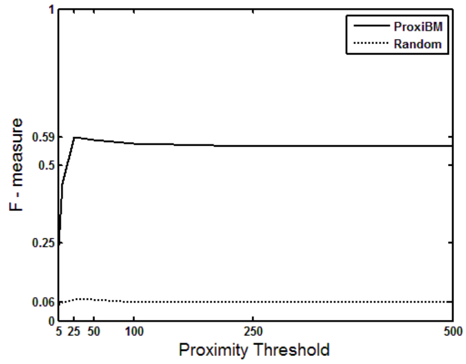

Figure 2: ProxiBM results vs. Random (Monte Carlo) results for different proximity threshold values.

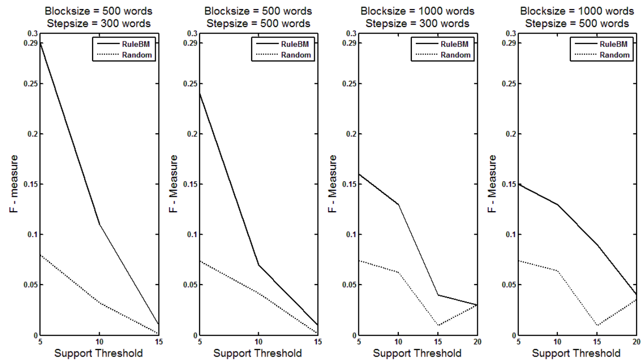

Figure 3: RuleBM results vs. Random (Monte Carlo) results for different blocksize, stepsize and support threshold values.

5.2 Social Network Analysis

In this section, we first show that the constructed social networks have the small world property. The social networks, which have small world concept, have short average path length and relatively high clustering coefficient [14]. In Table 2, we provide the network properties the average path length l and the clustering coefficient C of the social networks for ProxiBM and RuleBM with their best configuration and also for the manually created ground truth network. For comparison, we also provide clustering coefficient C_{rand} of the random graphs that have the same average degree and size of each network. The results show that the networks hold the small world property.

Table 2: Characteristics of social networks for both methods with the best configuration and ground truth compared to random networks that have the same average degrees $<k>$ and size (no. of nodes in network) values.

Network	*Size*	*No.of Edges*	$<k>$	l	C	C_{rand}
Ground Truth	71	321	9.04	1.90	0.93	0.15
ProxiBM	88	395	8.98	3.34	0.83	0.11
RuleBM	82	515	12.56	2.32	0.86	0.14

We also examine the social networks about node degree distribution. $P(k)$ is the distribution function and it denotes the probability that a randomly selected node has k edges. The distribution function of social networks with small world property has a power-law degree distribution whereas in the randomly generated networks with the same network properties it has a Poisson distribution [14]. Power law distribution is denoted as follows.

$$P(k) \sim k^{-\gamma} \tag{4}$$

In the above formula γ is called the scaling factor. Networks with power law distribution are called scale-free networks. In Power-law degree distribution, there are numerous nodes that have small degrees whereas few nodes have high degrees [14]. Degree distribution of the characters with the highest degrees for the actual network and the networks that are created by RuleBM and ProxiBM are listed in Table 3. The table shows that for decreasing degree k values, frequency of nodes with the same degrees increases. The degree distributions of all constructed networks for *Bitlis Section* follow a power law distribution rather than a Poisson distribution. The scaling factor γ of the distribution is calculated as 1.79 for actual network, 1.75 for ProxiBM and 1.71 for RuleBM best performing networks.

Table 3: Degree distribution of characters with the highest degrees for best configurations of ProxiBM and RuleBM along with actual social network for *Seyahâtname-Bitlis Section*.

Actual Network		ProxiBM		RuleBM	
Character	Degree	Character	Degree	Character	Degree
Abdâl Hân	69	Ziyâeddin Beğ	20	Abdâl Hân	66
Ziyâeddin Beğ	20	Selmân	20	Beşaret Ağa	40
Şeref Beğ	18	Beşaret Ağa	19	Şeref Beğ	32
Beşaret Ağa	17	Şemseddîn	16	Hüsrev Paşa	30
Haydar Ağa	15	Hasan Beğ	16	Haydar Ağa	27
Cündevân	13	Haydar Ağa	14	Zâl Paşa	27
Salmân-u Buhtî Ağa	13	Şeref Beğ	13	Salmân-u Buhtî Ağa	27
Racoy Ağa	13	Maktûl Haydar Kethudâ	13	Ziyâeddin Beğ	26
Bedir Beğ	13	Racoy Ağa	13	Şeref Beğ	23
Şemseddin Beğ	13	Seyfi Ağa	13	Âlemşâh Beğ	23
Âlemşâh Beğ	13	Bedir Beğ	13	Yaşar Beğ	23
Kerrârkulu Beğ	13	Siyâvuş	13	Cündevân	23
Yaşar Beğ	13	Kâzım Sührâb	13	Racoy Ağa	23
Seyfi Ağa	13	Salmân-u Buhtî Ağa	12	Seyfi Ağa	23
Vildân	12	Kevkeban	12	Süleymân Hân	23

6. CONCLUSION AND FUTURE WORK

In this work we introduce two methods for constructing social networks: ProxiBM and RuleBM by using textual data and apply it to the Bitlis Section of Seyahatnâme. We also check if the constructed networks hold social network properties. Both methods generate meaningful social networks which are significantly different from random and substantially similar to the social network manually constructed by a scholar-historian. The experimental results show that the networks created by ProxiBM show a higher similarity to the manually created social network than those of RuleBM. However, the disadvantage of ProxiBM is that it requires more focused (cohesive) blocks obtained from paragraphs. On the other hand, RuleBM is more flexible since it simply exploits blocks obtained from a sliding text window.

It is possible to obtain a better performance with RuleBM if we use contextually meaningful sliding text blocks: Our preliminary experiments with sliding paragraph-based text blocks provide evidence in that direction. For the construction of such cohesive units we may use an automatic text segmentation method [7]. Obtaining blocks automatically is also important for ProxiBM since natural cohesive units may not be readily available and their manual tagging is usually impractical. Furthermore, the results of these two methods can be combined in some elaborate data fusion techniques.

7. ACKNOWLEDGEMENTS.

This work is partially supported by the Scientific and Technical Research Council of Turkey (TÜBITAK) under the grant number 109E006. Any opinions, findings and conclusions or recommendations expressed in this article belong to the authors and do not necessarily reflect those of the sponsor.

8. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Database mining: A performance perspective. *IEEE Transactions on Knowledge and Data Engineering*, 5:914–925, 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, pages 487–499, San Francisco, CA, USA, 1994. Morgan Kaufmann Publishers Inc.

[3] S. Brin, R. Motwani, J. D. Ullman, and S. Tsur. Dynamic itemset counting and implication rules for market basket data. *SIGMOD Rec.*, 26:255–264, June 1997.

[4] R. Dankoff. *Evliyâ Çelebi in Bitlis: The relevant section of the Seyahatnâme*. E. J. Brill, Netherlands, 1990.

[5] P. Erdős and A. Rényi. On random graphs. I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[6] D. Hawking and P. Thistlewaite. Relevance weighting using distance between term occurrences. http://david-hawking.net/pubs/HawkingT96.pdf, 1996.

[7] M. A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 23(1):33–64, 1997.

[8] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice-Hall, 1988.

[9] S. A. Kahraman and Y. Dağlı. *Evliyâ Çelebi Seyahatnâmesi 4. Kitap*. Yapı Kredi Yayınları, İstanbul, 2003.

[10] C. Karbeyaz, E. F. Can, F. Can, and M. Kalpakli. A content-based social network study of Evliyâ Çelebi's *Seyahatnâme-Bitlis Section*. In E. Gelenbe, R. Lent, and G. Sakellari, editors, *Computer and Information Sciences II*, pages 271–275. Springer London, 2011.

[11] C. D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[12] J. Preiser-Kapeller. Calculating Byzantium? Social network analysis and complexity sciences as tools for the exploration of medieval social dynamics. http://www.oeaw.ac.at/byzanz/repository/Preiser_ WorkingPapers_ Calculating_ I.pdf, 2010.

[13] T. Raeder and N. V. Chawla. Modeling a store's product space as a social network. In *Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining*, pages 164–169, Washington, DC, USA, 2009. IEEE Computer Society.

[14] A. Özgür and H. Bingöl. Social network of co-occurrence in news articles. In C. Aykanat, T. Dayar, and I. Körpeoglu, editors, *Computer and Information Sciences - ISCIS 2004*, volume 3280 of *Lecture Notes in Computer Science*, pages 688–695. Springer Berlin Heidelberg, 2004.