Analyzing Persian Social Networks: An Empirical Study

Leila Esmaeili, University of Qom, Iran
Mahdi Nasiri, Iran University of Science and Technology, Iran
Behrouz Minaei-Bidgoli, Iran University of Science and Technology, Iran

ABSTRACT

Analysis of data in social networks is very important for researchers, sociologists, and academics. Given the size and diversity of web data in a Web 2.0 environment, analyzing this data has been a challenge. Since data act as inputs in such projects, the accuracy of the output is directly related to the input. Good data allows for extraction of valuable knowledge. In this article, the authors present their experiences with preparation and preprocessing of data in a Persian social network. The authors also report on the analysis of the data and findings.

INTRODUCTION

In recent years there has been a growing interest in web mining using scientific, social, political, and economic techniques. Organizations in different countries invest in social network analyses for different reasons. Analyses are mainly based on three methods: content analysis, structure analysis, and usage analysis. The type of analysis determines the manner in which researchers collect and prepare their data sets.

General web 2.0-based information systems, due to their free and interactive natures generate data that is not appropriate for web mining and knowledge discovery. A large amount of data generated is textual in nature and needs to be pre-processed before knowledge discovery. Moreover, systems change over time, and the amount and type of data generate changes. Therefore, unstructured data, which is not in the right format has to be transformed into structured, usable data that can be used in web mining studies. This process of conversion is complex and time consuming. In general, preprocessing techniques, if performed before data mining, can significantly improve the mining process, and reduce processing time.

In most web mining, data mining, text mining, and social network analysis studies and projects, data preprocessing is considered an important stage. But many of these studies...
are conducted without preprocessing due to the difficulty in collecting large amounts of data. Sometimes studies are conducted using smaller data sets to make preprocessing faster.

In our analysis of Persian social networks we did not come across any study that included complete preprocessing of data. Most of the research was conducted using blogs. Studies were based on data that had been collected by other researchers or in some cases Persian data was translated into English and then preprocessed (Sheykh Esmaili, Jamali, Neshati, Abolhassani, & Soltan-Zadeh, 2006; Sahebi, Oroumchian, & Khosravi, 2008). Esmaeili et al. (2011) used data stored in Parsi-yar Persian social network database to personalize recommended groups to users of the social network (Esmaeili, Nasiri, & Minaei-Bidgoli, 2011). The studied database consisted of content data and to some extent structured data. The data set was a raw one, which was analyzed for the first time. Complete preprocessing of a large volume of data in a Persian social network for the first time, classification of textual features are some of the study’s strengths. In this study, we elaborate on some preprocessing experiments and provide details of statistical and network analysis of the data set.

The data set employed in our attempt included data from a Persian social network called Parsi-yar. Parsi-yar contained activities for 5 years and 6 months for 78467 users, 3359 groups within 19 categories, and 275 groups without a specified category (Table 1). Data could be classified into three categories: user information, group information, and other information. The category, other information, included user interactions in the network, their public and private messages, users’ comments on messages, their friends, and their groups.

The organization of the rest of the paper is as follows. In the next section we explain the framework of data processing. This is followed by statistical analysis of research data and by network analysis of the data set. In the final section we provide a conclusion and areas of future research.

### Table 1. Subjective classification of groups (active and inactive groups)

<table>
<thead>
<tr>
<th>Category</th>
<th>Num. of groups</th>
<th>Category</th>
<th>Num. of groups</th>
<th>Category</th>
<th>Num. of groups</th>
</tr>
</thead>
<tbody>
<tr>
<td>Revolution</td>
<td>100</td>
<td>Sport</td>
<td>167</td>
<td>History</td>
<td>37</td>
</tr>
<tr>
<td>Social</td>
<td>716</td>
<td>Entertainment</td>
<td>332</td>
<td>Literature</td>
<td>117</td>
</tr>
<tr>
<td>Sciences</td>
<td>191</td>
<td>Game</td>
<td>75</td>
<td>Morality and spirituality</td>
<td>129</td>
</tr>
<tr>
<td>Youth</td>
<td>72</td>
<td>Familial</td>
<td>28</td>
<td>News</td>
<td>71</td>
</tr>
<tr>
<td>Geography</td>
<td>41</td>
<td>Art</td>
<td>182</td>
<td>Hygiene</td>
<td>30</td>
</tr>
<tr>
<td>Buy and sell</td>
<td>31</td>
<td>Religion</td>
<td>335</td>
<td>Computer</td>
<td>352</td>
</tr>
<tr>
<td>Business</td>
<td>78</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**DATA PREPROCESSING FRAMEWORK**

Users usually register for membership of social networks without providing complete information about themselves. They either volunteer data for required items or give worthless or meaningless information, and only a few of them correct this in subsequent visits. This has been attributed to factors such as privacy concerns, or the rush to register quickly. Thus, there is always the problem of lack of data or worthless data in our social network databases. Since databases usually contain data in different fields with different data types, each of these data types demands different preprocessing techniques. Determining the best preprocessing techniques for each category of data type plays an important role in improving the quality and
Related Content

Influence of Social Networks in the Decision to Vote: An Exploratory Survey on the Ecuadorian Electorate
[www.igi-global.com/article/influence-of-social-networks-in-the-decision-to-vote/139778?camid=4v1a](www.igi-global.com/article/influence-of-social-networks-in-the-decision-to-vote/139778?camid=4v1a)

Acceptance, Use, and Influence of Political Technologies among Youth Voters in the 2008 US Presidential Election
[www.igi-global.com/article/acceptance-use-influence-political-technologies/47197?camid=4v1a](www.igi-global.com/article/acceptance-use-influence-political-technologies/47197?camid=4v1a)

The Impact of Individual Differences on Social Communication Pattern in Online Learning
[www.igi-global.com/chapter/impact-individual-differences-social-communication/39808?camid=4v1a](www.igi-global.com/chapter/impact-individual-differences-social-communication/39808?camid=4v1a)
Guarding Corporate Data from Social Engineering Attacks
www.igi-global.com/chapter/guarding-corporate-data-social-engineering/39811?camid=4v1a