"Looking for a Supervisor": A Scientific Community Data Mining Application

Hao Xu[1, a], Zhuang Li[2, b] Jinchao Zhu[2, c], Albert Bellman[3, 1, d] Yunfeng Wei[1, e], Dingke Song[2, f] and Lan Huang[2, g*]

[1]College of Computer Science and Technology, Jilin University, Changchun 130012, China;

[2]College of Software, Jilin University, Changchun 130012, China;

[3]School of Computing, University of Eastern Finland, Joensuu 80110, Finland;

[a]xuhao@jlu.edu.cn, [b]lizhuang14@mails.jlu.edu.cn, [c]zhujc14@mails.jlu.edu.cn
[d]albert.bellman@cs.uef.fi, [e]weiyf@mails.jlu.edu.cn, [f]songdinke@mails.jlu.edu.cn,
[g]huanglan@jlu.edu.cn

*Corresponding Author

Keywords: Data Mining, Scientific Publications, Clustering

Abstract. We all live in the information society. Social networks in university campuses produce linked information. However, these social networks are not targeted to the relationship between students and teachers, neither easily established. As a consequence, students don't have a platform system where to look for a supervisor. In this paper, we present an application called "Looking for a Supervisor" that uses Python and R for data mining purposes. The front end web design is developed with PHP. This system has attained a very good communication environment between teachers and students. It also benefits the students on the facilitation of their research topics of interest.

Introduction

Students trying to find a supervisor are mostly looking for information provided by some websites. However, students need to screen many pages of information regardless of knowing how up to date this information is. It would be more effective for students to find a supervisor in line with their own interest of research direction if there is an intelligent system to gather and analyse available information. In turn, the candidate supervisor cannot select their students according to the characteristics of each student. A bridge of mutual understanding between students and teachers is missing [1].

In this paper, we show the design and implementation of an application called "Looking for a Supervisor". The students can search for their future corresponding supervisor according to different interests and research direction. They can easily see each supervisor's published papers, research topics and other related data. At the same time, the students' own profile information will be displayed for teacher reference. In addition, students and teachers can communicate online to enhance their mutual understanding and common interests [2]. The system can determine the recent research direction according to recent publications indexed by Ei Compendex and Web of Science. The method of data mining supports the whole system at the bottom layer. Firstly, we obtain the latest data from databases, including authors and summary of the paper. Right after, the system clusters the authors studying in related fields. Its purpose is to ensure the reliability and real-time basis of the information [3].

System Design

The entire application framework is built-in PHP. A MySQL database stores the title, supervisor's information, summaries and additional data [4]. The primary analysis tool s use Python and R. The main task of Python is web crawling in order to obtain the required data. R is used for data analysis and for the provision of charts and graphs.

A. Main implementation process of the system. Firstly, the application judges whether a student or a teacher logs in the system. Students can input the keyword of research topics to find a supervisor. The system will match qualified supervisors and display them on the interface. A representation of the system flow chart is displayed in Figure 1.

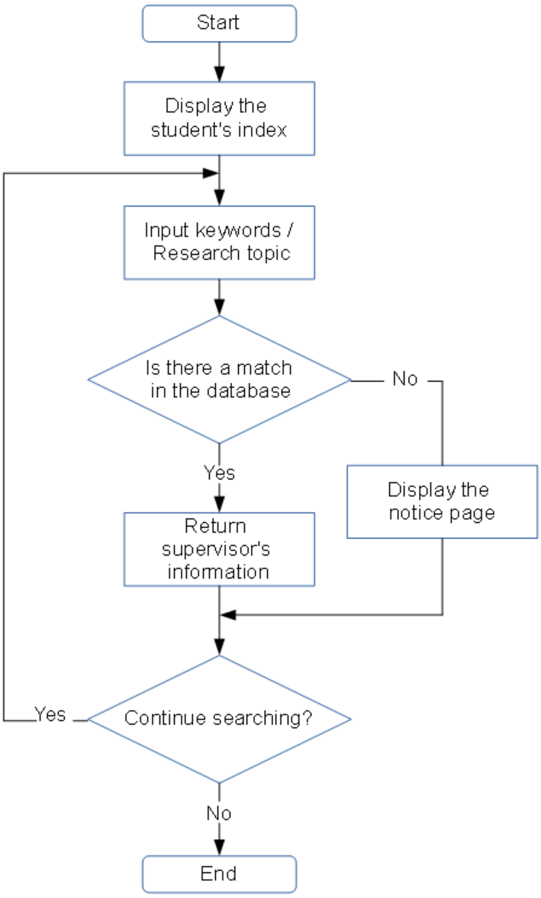
Start
Display the student's index
Input keywords / Research topic
Is there a match in the database
No
Yes
Display the notice page
Return supervisor's information
Yes
Continue searching?
No
End

Fig. 1 System Flow Chart

B. Data acquisition. The first step in data analysis is the data acquisition. The web crawler can search the paper across databases. The system was designed to search for the last three years papers' at the College of Computer Science and Technology of Jilin University. It gathers each papers' authors, titles and abstract by iteration. Due to the large number of results attained, we use multiple threads to improve data crawling speed. All this data is stored in the structured MySQL database [5].

C. Data pre-processing. Amongst the acquired data, certain names may appear using varied characters in different papers. In these particular cases, we need to make the following pretreatment for the author's name:

1. Word segmentation for each paper's author name.
2. Convert author names into lowercase.
3. Remove symbolic characters in the middle of some author's name.

After the above pretreatment process, the author's name is consistent with an appropriate format for data mining processing.

D. Data mining. Using the R programming language, we can place each of the two co-authored author in a tuple. For example, a paper has six authors, totaling C62 tuples. Using R's drawing package, we are able to draw a correlation of all authors in the database. After obtaining co-author relationship, the distance between each supervisor is generated automatically [6].

We use classic K-means clustering algorithm of data mining to cluster supervisors and generate the research community. Finally, we calculate the abstracts' word frequency of each class by using R in order to get the research directions of each research community [7,8]. The resulting cluster graph is shown in Figure 2.

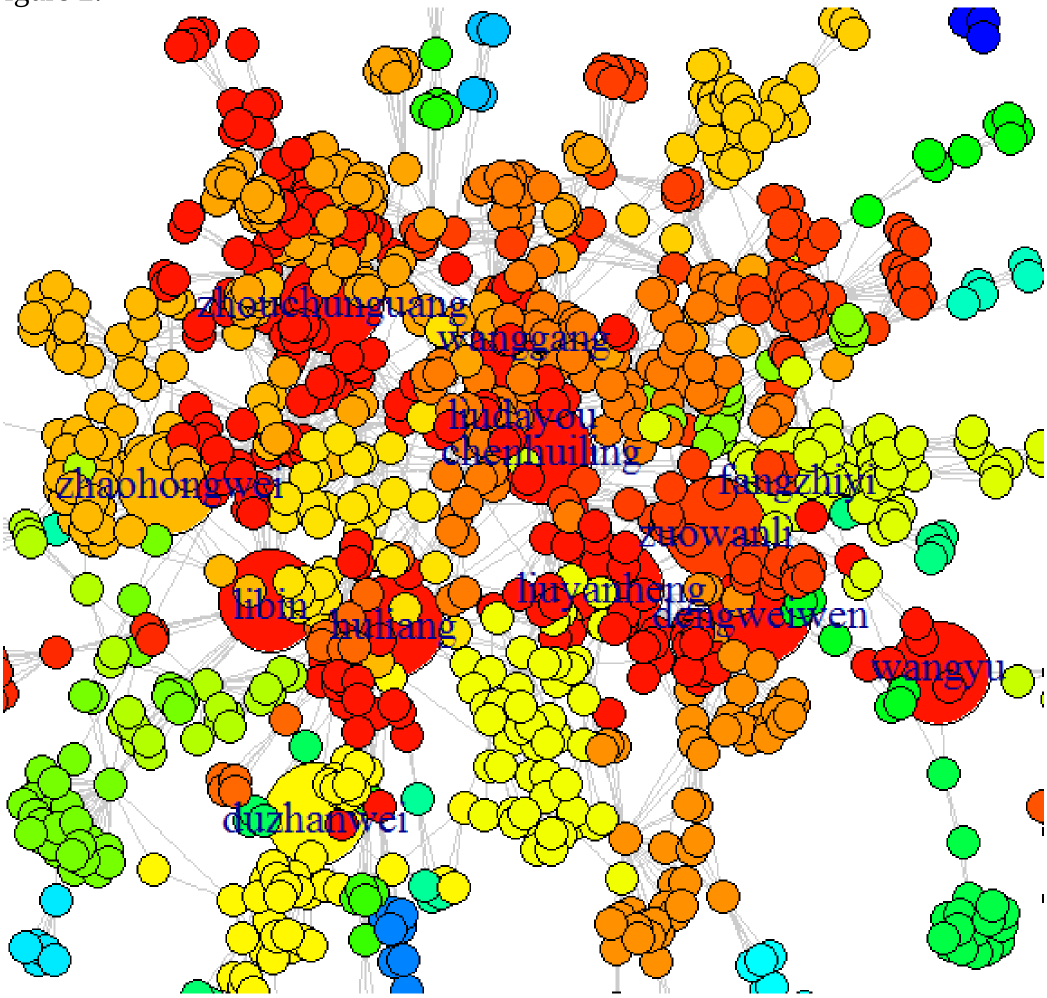

Fig. 2 Clustering by co-authorship

Conclusion and future work

In this paper, we have shown the design and implementation of a data mining system in order to find an adequate supervisor through its publications. It is convenient for students and future supervisors, providing an aid to their common interests. The application has several opportunities for improvement. For example, if names are expressed through different abbreviations the system is not able to recognize such differences yet. Our future work is to integrate a database including additional sources of information, such as their professional activities, conference attendances and presentations in scientific venues [9]. The application will also possess the added possibility to automatically recommend a qualified supervisor for the students.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.61300147,No.61472159), China Postdoctoral Science Foundation (No.2014M551185), Electronic Commerce Engineering Laboratory Project of Jilin Province (2014N143), the Science and Technology Program of Changchun (No.14GH014) and the Erasmus Mundus Swap and Transfer Project of the EU Commission.

References

[1] Hao Xu, Chang-hai Zhang, Yu-an Tan, Jun Lu: An improved evolutionary approach to the Extended Capacitated Arc Routing Problem. Expert Systems with Applications. 38(4): 4637-4641, 2011

[2] Ahmad Shaker Abdalrada Alkunany Thaer Farag Ali, Manage Website Template That Using Content Management System Joomla, Journal Of Wassit For Science & Medicine, 2013, pp.152-161

[3] Hao Xu, Bo Yu: Automatic Thesaurus Construction for Spam Filtering Using Revised Back Propagation Neural Network. Expert Systems with Applications. 37(1): 18-23, 2010, pp.

[4] Santos, J., Mendonca, J., Martins, J.C. Instrumentation remote control through internet with PHP[C]. Virtual Environments, Human-Computer Interfaces and Measurement Systems, 2008. VECIMS 2008. IEEE Conference on,2008,6.

[5] Hao Xu, Fausto Giunchiglia. SKO Types: An Entity-based Scientific Knowledge Objects Metadata Schema. Journal of Knowledge Management, Volume 19, Issue 1, 2015

[6] Karout Salah A, Gdeisat Munther A, Burton David R et al. Two-dimensional phase unwrapping using a hybrid genetic algorithm.[J]. Applied Optics, 2007, 46(5).

[7] Hao Xu, Yue Zhao, Li-ning Xing, The Novel Heuristic for Data Transmission Dynamic Scheduling Problems. Journal of Applied Mathematics, Volume 2013.

[8] Pyle, D. Data Preparation for Data Mining [M]. Morgan Kaufmann, 1999.

[9] Christos Mettouris; George A. Papadopoulos, Ubiquitous recommender systems, Computing, 2014, pp.223-257.