Combining Web Data Extraction and Data Mining Techniques to Discover Knowledge

Abstract

- The design and implementation of a support system for Knowledge Discovery is the challenge of many researchers.
- As Data Mining is the main key step in Knowledge Discovery process in Databases (KDD), it is necessary to finda new methodology that combines web data extraction playing the role of data collection from the web and data mining techniques on the extracted categorical data in order to discover knowledge.
- The main contribution of this research is proposing a methodology to apply the clustering notion on categorical web data and to use the clustering results as part of the input for the classification conducted on another set of data.
- Data mining and relative data processing are conducted by developing intelligent tools.
- The performance of the algorithms used in our methodology is demonstrated with the clustered job postings dataset and classified job searchers dataset by using the three measures accuracy, recall and precision for the clustering algorithm and the error of classification for the classification technique.

Existing system

- As Data Mining is the main key step in Knowledge Discovery process in Databases (KDD), it is necessary to find a new methodology that combines web data extraction playing the role of data collection from the web and data mining techniques on the extracted categorical data in order to discover knowledge.

- In labor market, these obscurities are becoming more and more challenging for job searchers, employers and recruiting agencies, aiming all together to take advantage of new ways of recruitment.
- Job searchers are navigating the large number of job postings advertised on multiple job board websites.
- There is an extensive need to provide job searchers with an interface that integrates and analyzes job postings based on given attributes such as job sector, salary range and required experience.

Disadvantages

- Distributed data over millions of different web servers.
- Volatile data: Many web documents disappear rapidly.
- Large volume: Billions of separate documents exist on the web.
- Unstructured data: No uniform structure for web documents.
- Redundant data: A lot of duplicate documents exist on the web.
- Data quality: As there is no editorial control for web documents, the quality of writing is poor.
- Heterogeneous datapresented as structured tables, texts, images, multiple media types and other types.
- Hidden patterns: Patterns used to understand existing data and to predict how new instances will behave are missing.
- Segmentation problems: There is a need to assign large volume of data into a relatively small set of groups.

Proposed system

- main contribution of our proposed approach consists of a methodology that combines web data extraction and data mining techniques in order to discover knowledge.The main goals associated with the recruitment needs are enumerated as follows:
- 1) Applying a suitable embedded tool for web data extraction to extract data (job postings) from several recruitment web sites.
- 2) Developing an intelligent tool “Jobs Mining” that aids in processing and consolidating extracted job postings from several sources in order to end up with the dataset 1, clustering the categorical data in dataset 1 into a relatively small set of groups, classifying another dataset 2 (job searchers) in such a way to achieve predicting how new instances will behave, and assigning right job postings to right job searchers based on the results of the clustering and the classification.
- 3) Experimenting the results based on accuracy, precision and recall for the Clustering technique and error of classification for the Classification technique.
- 4) Deploying the discovered knowledge and results by sending automatic emails to job searchers informing them about job opportunities that fit their needs.

Advantages

- The results show that our proposed approach of combination ends up with good results in Knowledge Discovery from the web
- identified characteristics of the most useful data mining techniques and developed two algorithms, k-mode clustering and Naïve Bayesian classification, that can be used to predict useful Fields
- The experimental results show that the accuracy of the clustering algorithm is 92.53%

HARDWARE REQUIREMENTS

- Processor – Pentium –III
- Speed – 1.1 Ghz
- RAM – 256 MB(min)
- Hard Disk – 20 GB
- Floppy Drive – 1.44 MB
- Key Board – Standard Windows Keyboard
- Mouse – Two or Three Button Mouse
- Monitor – SVGA

SOFTWARE REQUIREMENTS

- Operating System : Windows 8
- Front End : Java /DOTNET
- Database : Mysql/HEIDISQL

Conclusion

- In this paper, we firstly used an embedded tool for web data extraction to collect job postings, secondly, we identified characteristics of the most useful data mining techniques and developed two algorithms, k-mode clustering and Naïve Bayesian classification, that can be used to predict useful fields.
- We ended up with two meaningful clusters of job postings with fair distribution (62% for C1 and 38% for C2) in which the cluster 2 is matching to all the job postings requiring high qualifications.
- The experimental results show that the accuracy of the clustering algorithm is 92.53%. As for the classification, we deduced that the classification error decreases with the increase of the training set. For 75% training set with 25% validation.
- we found the classification error 4.9% which is satisfactory. Despite the absence of results comparison with previous works adopting the same methodology, we consider that these results are good.

Future work

- As this case study contributes in promoting employability, concerned parties could take advantage of such tools to support people to find jobs with the collaboration of recruitment agencies.
- In addition to providing a recommended tool for job searchers, our approach could be developed in the future to contribute for gaining insights on the required skills and the distribution of jobs across the sectors and countries in the region.
- Furthermore, with the absence of a valid occupations list at the national level, such methodologies could be adopted to conduct studies on the labor market needs and take proper actions for preparing future employees to fulfill these needs

Reference

[1] Ng, M. K., Li, M. J., Huang, J. Z., & He, Z. (2007). On the impact of dissimilarity measure in k-modes clustering algorithm. IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(3).

[2] Huang, Z. (1997). A fast clustering algorithm to cluster very large categorical data sets in data mining. DMKD, 3(8), 34-39.

[3] Chang, C. H., Kayed, M., Girgis, M. R., & Shaalan, K. F. (2006). A survey of web information extraction systems. IEEE transactions on knowledge and data engineering, 18(10), 1411-1428.

[4] Mohaghegh, S. D. (2003). Essential Components of an Integrated Data Mining Tool for the Oil & Gas Industry, With an Example Application. In in the DJ Basin. Paper SPE 84441 presented at the SPE Annual Technical Conference and Exhibition.

[5] Tan, P. N. (2006). Introduction to data mining. Pearson Education India.