Chapter 92
New Regional Investors Discovery by Web Mining

Ting Chen, Jian He and Quanyin Zhu

Abstract In order to promote micro business and sell their products and services, a new proposed system by web mining technology is used to discover new regional investment project. Project information published on the government websites of Huaian, Jiangsu province is extracted to utilize the proposed method. Python language, MySQL database and Django web framework are used to develop the application system, and the multi-factor matching algorithm is provided to collect key information of the project name, time, address, contact, domain and URL by web mining. Furthermore, the location and statistics functions are accomplished in the proposed system which can meet application requirements of micro business.

Keywords Investors information discovery · Micro business · Web mining · Python language · MySQL database · Django web framework

92.1 Introduction

In recent years, the rapid development of Internet brings geometric growth to web information, and the vast volume of information means it has the characteristic of polynary and redundant as well. Web pages cannot be directly made use of by traditional database systems for its semi-structured characteristic [1]. How to use the information better becomes the focus of attention. At present, most web pages are given to the semi-structured document in HTML form. Web documents can be represented as unstructured documents, semi-structured document and structured document. For the correct extraction of web information, a lot of work has been done at home and abroad. References [2] proposed that results of participle

T. Chen (✉) · J. He · Q. Zhu
Faculty of Computer Engineering, Huaiyin Institute of Technology,
Huaian, China
e-mail: apple_ting@126.com

W. E. Wong and T. Ma (eds.), *Emerging Technologies for Information Systems, Computing, and Management*, Lecture Notes in Electrical Engineering 236, DOI: 10.1007/978-1-4614-7010-6_92,

algorithm could touch the shopkeepers' minds, and it can support the originality data for the commodities markets and dynamic trend analysis. Reference [3] describe the new API for data mining proposed by Microsoft as extensions to OLE DB standard. Reference [4] describe data mining and brainstorms on its application to power systems. Reference [5] supported some actions to promote information and communications technologies in the autonomous community of the region of Murcia of Spain. Reference [6] analyzed the impact of urbanization on regional flood risk in the Qinhuai River Basin of China. Reference [7] applied research of the ash connection analysis in highway's influence of regional economies. But all of them not reported the useful method to discover new regional investment project using web mining technologies. So our proposed is how to find those new regional investment project and help the modern micro business to solve their concerned at the recent development.

Based on our past work [8–12], we select investor's discovery and web mining pages from the well-known colleges and universities in China in order to build an efficient system of new regional investor's discovery by web mining and study the Multi-factor matching method for information ex-traction. The second part gives the system architecture for new regional investor's discovery. In the third part we introduced web crawler strategy of webpage search. Multi-factor matching algorithm is proposed in the fourth part. In the last two parts we gave application system structure and result of system running.

92.2 System Design

92.2.1 System Framework and Development of Language

How to seek for specific investment projects investors from the web, mining data, and store in the database is serious problem for micro businesses. Fast and accurately discovering relevant project information and contacting information is very conducive for micro businesses to publicize and sell their products or services. To build a data mining system, the system should conclude the following functions:

1. Automatically mining the release date, URL and related data information. Such as, project overview, project introduction, project approval document and so on.
2. According to their own requirements and keyword, querying the data in the database easily through software interface.
3. Study of the basic knowledge of Python language.
4. Study of re-module in the regular expression matching.
5. Study of Django web development framework.
6. Study of MySQL data construction and query.

Django's advantage is simply and rapidly developed database-driven website. It emphasizes the importance of code reuse. Multiple components can be very convenient serving for the whole frame as plug-in form. Django has many powerful the third party plug-in. It makes Django Strong expansibility. It also stressed the rapid development and the principle of DRY.

With Python class form defining data model, ORM related model and database together. You will get a database API using very easily, and also you can use the original SQL statement in the Django language.URL assignment use regular expression matching URL. You can design URL random without specific limited framework.

Using Django powerful and extensible template language, the template system can be separated from design, content and Python code. And it also has inheritance.

92.2.2 Web Crawler

Web crawler crawls web information automatically following certain rules. Webpage search strategy can be divided into depth-first, breadth-first and best-first. Depth-first in many cases will lead to crawler trapped problems. And breadth-first and best-first are popular methods.

Breadth-first search strategy means that next level search must be after the completion of the current level in the crawl process. The algorithm design and implementation is relatively simple. Algorithm flow is shown as Fig. 92.1.

Best-first search strategy according to the web analytics algorithms to predict the similarity of the candidate URL and landing page, or correlation with the theme, and select the best rated one or several URLs to crawl. The only access through web analysis algorithm predicted "useful" pages. One problem is that

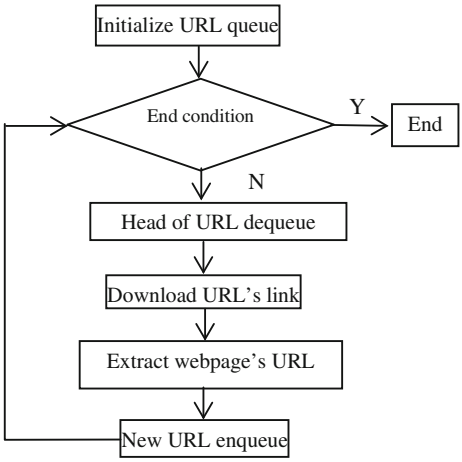

Fig. 92.1 Breadth-first search algorithm flow

reptiles crawl path on many relevant pages may be ignored, because the best-first strategy is a local optimum search algorithm. Therefore it needs to be combined with the best-first specific application to improve, to jump out of local minima.

92.2.3 Multi-Factor Matching Algorithm

This paper focuses on the example of chart showing for science expert information extraction.

Let the view-source of webpage be defined as D, body of the page be defined as S:

$$S \subset D \tag{92.1}$$

Let the normalized text be defined as $\hat{S}$:

$$\hat{S} \subseteq S \tag{92.2}$$

Keywords corpus consists of some self-learned keyword. Let the keywords corpus be defined as F, then F be presented as

$$F = \{f_1, f_2, \ldots, f_n\} \tag{92.3}$$

In fact, the expert information field included in F. Let it defined as f_n, the position of f_n in $\hat{S}$ is defined as k:

$$k = (f_n, \hat{S}) \tag{92.4}$$

Let the potential expert information field be defined as t, define a constant as con then t is equals to the normalized text range between k and $k + con$:

$$t = \hat{S}(k, k + con) \tag{92.5}$$

Finally, because of other fields affect, we must remove the affect. The positions of F except f_n is defined as K:

$$K = \{k_1, k_2, \ldots, k_n\} \tag{92.6}$$

Let define the min position in K as $k_{\min}$, define the real expert information as t_{real}, we can conclude that the real expert information filed:

$$t_{real} = t(0, k_{\min}) \tag{92.7}$$

92.2.4 Data Table

System data sheet is mainly used for storage government network project information and credit information publicly shared on all project information in the column. And it could be called and inquired by the system (Table 92.1).

92.2.5 System Data Structure

Application system design includes eight sheets, respectively as the auth_message sheet, django_conten_type sheet, auth_user sheet, auth_permission sheet, auth_user_user_permission sheet, auth_group_permission sheet, auth_user_group sheet and auth_group sheet. Its structure is shown in Fig. 92.2.

The data structure of application system is succinct than the other system. Our aim is adequate to the boss of micro business for application requirements. If using the web service technology, we can get the encapsulation web service [11, 12] for other application system and get the advantage of platform irrespective.

92.3 System Implementations

Each hyperlink function in the interface of the application system as follows: URL: Mining the demand URL in the webpage, and store in the URL sheet of the database. Data: Release date of mining project, and store in the date sheet of the database. Info: mining project summary, approval documents, posting date, and store in the Huaian sheet of the database. Query: query the project data and URL in the database. Project Analysis: Building materials can be divided into structural materials, decoration materials and some special materials. Structural materials include wood, bamboo, stone, cement, concrete, metal, brick, ceramics, glass, engineering plastics, composite materials, etc. Decorative materials include a variety of coatings, paints, coatings, veneer, colored tile, glass and other special effects. Special material includes waterproof, moisture-proof, anti-corrosion, fire,

Table 92.1 System data table

Name	Type	Length
id	int	10
doc	varchar	500
proview	varchar	1000
date	varchar	20
url	varchar	50
date2	varchar	50

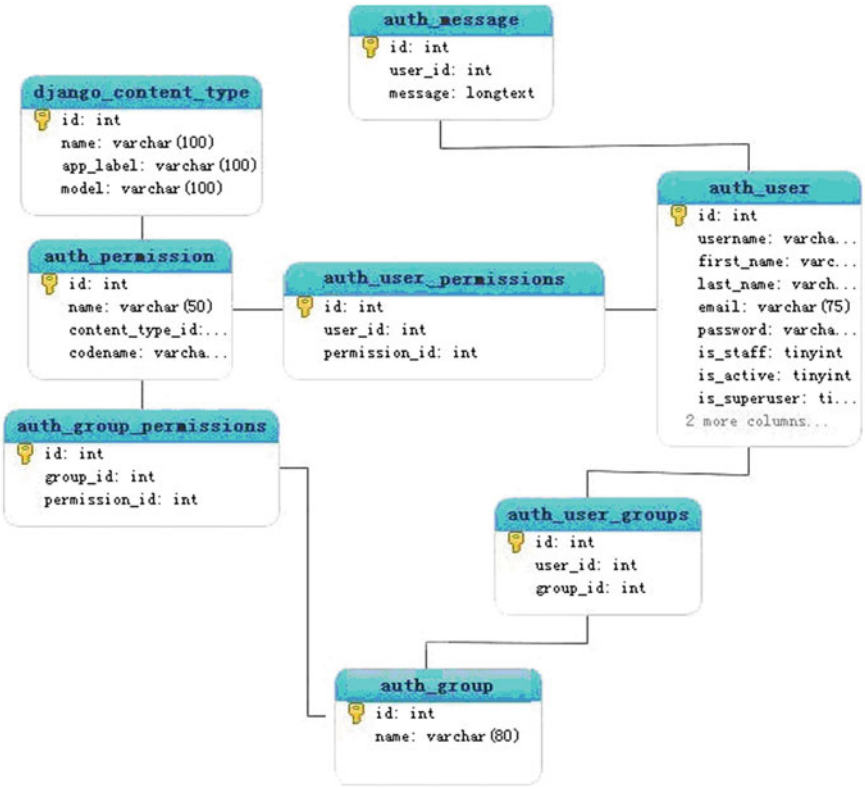

Fig. 92.2 Data structure of the application system

flame retardant, sound insulation, heat insulation, thermal insulation, sealing and so on. Figure 92.3 illustrates the connection information in a part of the URL database.

With the using of software, micro business based on building materials can find associated with their investment projects, promote and sell their products and services, thus derive profits to promote their own development.

Take Huaian of Jiangsu province for example, with the investment project comparison of nine counties in Huaian, micro business get an overall Huaian investment regional distribution map. With it, micro business promotes and market their product or service through its own geographical advantages. Figure 92.4 shows the investment project numbers of each region.

Take the industry for example, culture construction is the development of education, science, literature and art, the press and publishing, radio, television, sports and public health, library, museums and other cultural undertakings of activities. Micro business related with the cultural construction, find suitable for their own projects according to the search results, thus promote and sell their

id	url2
1	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=15257
2	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=15255
3	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=15260
4	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=15256
5	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=15101
6	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14747
7	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14694
8	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14418
9	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14051
10	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14112
11	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14116
12	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14120
13	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14108
14	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14106
15	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14100
16	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14098
17	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14118
18	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14104
19	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14053
20	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14102
21	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14122
22	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14114
23	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14110
24	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=14124
25	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=7247
26	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=6879
27	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=6824
28	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=6629
29	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=6569
30	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=6570
31	http://222.184.59.60:8080/gcjsly/content/cont_xmspgk.jsp?styleName=sjzf&articleId=6571

Fig. 92.3 Portion of URL extracted from the database

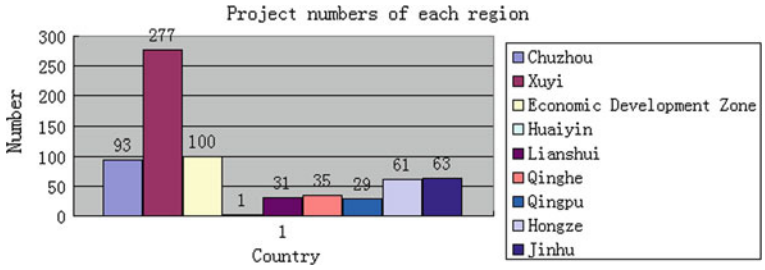

Fig. 92.4 Statistical data of nine counties

products or services. Figure 92.5 shows the investment project status according to the industry classification.

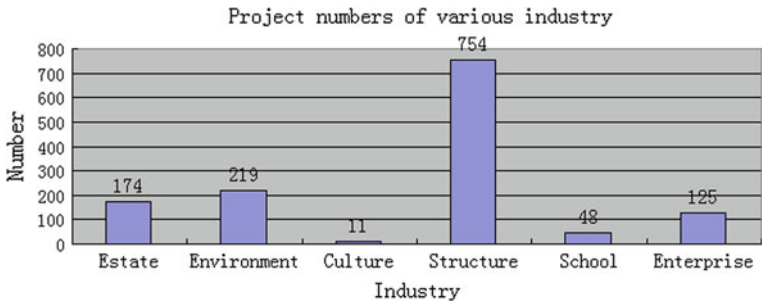

Fig. 92.5 The investment project status of various industry

92.4 Conclusion

Seeking for specific investment projects investors from the web, mining data, and storing in the database are serious problems for micro businesses. With the technology of data mining, discovering the new regional investment projects targeted the promotion of enterprise products or services can reduce the operation cost of small and micro businesses. The proposed systems can satisfy small and micro businesses for informatization and can meet market requirements.

Python language, MySQL database and Django web framework are used to develop the proposed system which can extract the region's investment project information from government websites. All the information of investment project can be stored in the database.

References

1. Li, Y.G., Sun, H.Y., Lin, S., et al.: Web information extraction based on hidden Markov model. In: Proceedings of the14th International Conference on Computer Supported Cooperative Work in Design, pp. 234–238 (2010)
2. Hoffman, P., Grinstein, G., Marx, K., et al.: DNA visual and analytic data mining. In: Procedings of the Visualization 1997, pp. 437–441 (1997)
3. Netz, A., Chaudhuri, S., Fayyad, U., et al.: Integrating data mining with SQL databases: OLE DB for data mining. In: Proceedings of 17th International Conference on Data Engineering, 2001, pp. 379–387 (2001)
4. Madan, S., Won-Kuk, S., Bollinge, K.E.: Applications of data mining for power systems. In: Proceedings of the IEEE 1997 Canadian Conference on Electrical and Computer Engineering, pp. 403–406 (1997)
5. Escudero-Sanchez, M., Pavn-Mario, P., Fernandez-Caceres, J.L.: Some actions to promote information and communications technologies in the autonomous community of the region of Murcia (Spain). In: Proceedings of the 11th Mediterranean Electrotechnical Conference, pp. 163–167 (2002)
6. Shi, Y., Xu, Y.P., Cai, J.: Analysis of the impact of urbanization on regional flood risk: A case study in the Qinhuai River Basin, China. In: Proceedings of the 19th International Conference on Geoinformatics, pp. 1–6 (2011)

7. Zhao, Q.W., An, Y.H.: The applied research of the ash connection analysis in highway's influence of regional economies. In: Proceedings of the Seventh International Conference on Fuzzy Systems and Knowledge Discovery 2010, pp. 1073–1076 (2010)
8. Zhu, Q.Y., Zhou, P., Cao, S.Q., et al.: A novel RDB-SW approach for commodities price dynamic trend analysis based on web extracting. J. Digit. Inf. Manag. **10**(4), 230–235 (2012)
9. Zhu, Q.Y., Yan, Y.Y., Ding, J., et al.: The case study for price extracting of mobile phone sell online. ICSESS 2011, pp. 282–285 (2011)
10. Ding, J., Zhu, Q.Y., Zhou, L.J., et al.: Research on the new products discovery based on web mining. MINES 2011, pp. 528–532 (2011)
11. Ding, J., Wu, B., Ding, T.T., Zhu, Q.Y.: The case study on service encapsulation for web-based application system. CSSS 2012, pp. 2684–2687 (2012)
12. Zhu, Q.Y., Zhou, H.Y., Yan, Y.Y., et al.: Research on the service encapsulation for web-based system. ICCDA **2011**(4), 535–538 (2011)