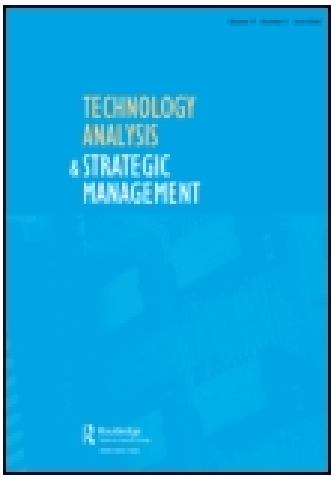

Technology Analysis & Strategic Management

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ctas20

Profiling science and innovation policy by object-based computing

Zhixiong Zhang[a], Jianhua Liu[a], Yimin Zou[b], Jing Xie[a] & Li Qian[a]

[a] National Science Library, Chinese Academy of Sciences, Beijing, China

[b] College of Economics and Management, Zhejiang Normal University, Jinhua, Zhejiang 321004, China

Published online: 05 Feb 2014.

Technology Analysis & Strategic Management, 2014
Vol. 26, No. 5, 581–593, http://dx.doi.org/10.1080/09537325.2014.881992

Profiling science and innovation policy by object-based computing

Zhixiong Zhang[a]*, Jianhua Liu[a], Yimin Zou[b], Jing Xie[a], Li Qian[a]

[a]*National Science Library, Chinese Academy of Sciences, Beijing, China;* [b]*College of Economics and Management, Zhejiang Normal University, Jinhua, Zhejiang 321004, China*

Named entities, such as key initiatives, research programmes, scientific strategies and policies, are research objects or objects that are embedded in many web pages of science and innovation institutes. These objects provide important information that can be extracted intelligently from those pages. This paper brings forward an object-based computing method by using objects and their semantic information for profiling science and innovation policies. After extracting the objects and the relationships between them, we proposed an object grid to represent web pages. Objects were transferred into machine-readable knowledge units with rich semantic information. By using computational objects, we judged the intelligence value of web resources and classified policies into more detailed categories, such as strategic plan, research and development, and budget. To test the effectiveness of profiling science and innovation policies by using the object-based computing method, this paper conducted a research profiling experiment system.

Keywords: research profiling; science and innovation policy; research objects; object grid; object-based computing

1. Introduction

Research profiling (RP) is an analysis method based on bibliometrics and text extraction tools. These tools could be used to broadly scan the contextual literature information and depict the research context and research efforts wisely (Johanna et al. 2007; Porter, Kongthon and Lu 2002). RP tools help researchers easily identify related topics within the research domain, sketch out a 'big picture' perspective on research activities, understand the research community, acquire insight on how innovation is progressing, and map (graphically represent) topical interrelationships of an entire research area.

Much work has been carried out on RP (e.g. Porter and Newman 2011; Choi, Le and Sung 2011; Guo, Huang and Porter 2010; Johanna and Jan 2007; Hicks and Atlanta 2004). The data they used for analysis were mainly from science, technology and innovation databases, such as Web of Science and Chinese Science Citation Database. Few researchers utilise web resources for RP or adopt RP tools, especially for web resources. Compared with high-quality formal databases, web

*Corresponding author. Email: zhangzhx@mail.las.ac.cn

resources are informal, unstructured, poor in semantic meaning, and occasionally unreliable and unstable. However, web resources still have the advantage in terms of size, timeliness, accessibility and diversity. With the help of semantic web mining technologies (Gerd, Andreas, and Bettina 2006), we could apply the profiling application into web resources. Sehgal (2007) proposed a profile-based approach to explore the social networks of US senators generated from web data. The social networks were then compared with networks generated from voting data. Despite these studies, no in-depth study on web content analysis is available.

The National Science Library (NSL) is an information analysis service provider of the Chinese Academy of Sciences (CAS). Learning timely science and innovation policies and providing research reports to policy makers are very important tasks for the library. Aside from the science, technology and innovation databases, some web pages released by key science and innovation related institutes, such as the Office of Science and Technology Policy of the White House, are also important resources. Several projects have been carried out in NSL to automatically or semi-automatically support the monitoring and profiling of the science and innovation policies released on the websites of those key institutes. In this paper, we focus on the web resources of some key institutes to implement science and innovation policy RP. We have then attempted to monitor and depict the development of science and innovation policies.

Based on science, technology and innovation databases, our project is similar to RP processes that require to deal with web page mining and content extraction, analysis and visualisation. Considering the great challenges to implement RP on web resources, the authors bring forward a new approach of object-based computing to facilitate the profiling of science and innovation policies based on web resources. The rest of this paper is structured as follows. Section 2 describes the main idea of object-based computing. Section 3 discusses some key issues of object-based computing for profiling. Section 4 shows some applications of the new approach. Section 5 concludes the paper and provides recommendations for future work.

2. Main ideas of object-based computing

To present the main ideas of object-based computing, the meaning of *object* should first be explained. The named entities embedded in the web page usually contain the core information of that web page. Such entities include scientific strategies, policies, key initiatives, research programmes and key research institutes. This content is valuable for the automatic mining of intelligence from web pages. For example, in a news article titled 'President Obama Lays out Strategy for American Innovation' (Obama 2009), two important named entities, namely, 'President Obama' and 'Strategy for American Innovation', are found. These two entities could present the main topics and value of the news directly. Thus, we consider these meaningful named entities as research objects or objects. Utilisation of these objects (meaningful named entities) and their inner relationships could further accomplish the computation for knowledge discovery. Therefore, an object-based computation method is used to identify the objects and the relationships between them from web pages to determine profiling science and innovation policies. Four main ideas are identified in this method:

- Web pages on science and innovation policies from the websites of key institutes, such as the Office of Science and Technology Policy and US National Science Foundation, should be crawled continuously to monitor the changes in science communities. The institutions should be chosen correctly in view of the authority and reliability of web resources obtained. The experts in the fields would be involved in the institute filter process.

- The crawled free texts should be transformed into time-stamped objects (when objects occur, that is, objects with temporal features) through information extraction. Important objects should be identified to support object-based computation. After related web pages are crawled from the targeted institutes' website, we need to transfer the unstructured free texts into structured and machine-readable object collections. Given the dynamicness of web resources, we should add the temporal feature of objects when labelling the extracted objects. Take 'July 13, 2010, White House Announces National HIV/AIDS Strategy' as an example. We turn this title into the following time-stamped triple: (object type, object value, time stamp) (Strategy, National HIV/AIDS Strategy, July 13, 2010); (Object A, Object B, Relationship Type, Time Stamp) (White House, National HIV/AIDS Strategy, Announces, July 13, 2010). Furthermore, we should record other context information of the objects, such as source, syntax and position in the text. With the information of objects extracted from the text, we could construct an object grid to identify the important objects, which will be explained in detail in the following section.
- A large-scale knowledge base based on time-stamped objects should be built to achieve semantic mining of the related topics. The large-scale knowledge base with time-stamped objects is important for object-based computation. In our method, all extracted objects and their semantic information are preserved in relation databases. These objects and their relationship representation model can be extended to content analysis tasks with a temporal dimension, such as burst topic detection (Kleinberg 2002).
- The status of science communities should be profiled by using information visualisation. Basing on enormous objects, we identify new science and innovation policies, assess the intelligence value of those policies to support Chinese scientific policy-makers, classify the policies into more detailed categories (such as formal scientific declaration, strategic plan, research and development, budget and organisation restructure), outline the hot topics in a period of time, depict current active activities and players in the institutes, and cluster the related policies of the institutes.

Given that web crawling and information visualisation are not the key issues of object-based computation, this paper specifically focuses on object identification and object-based computation.

3. Key issues

As mentioned above, object identification and object-based computation are the key issues of this paper. First, we discuss identification of important objects.

3.1. *Identification of important objects*

As shown in Section 2, the named entities embedded in web pages usually carry the core information of certain web pages. These named entities could present the main topics and value of the news directly. Extracting these entities from the unstructured web pages is important. An object grid transfers the objects embedded in the web into structured and machine-readable knowledge units. An object grid is a two-dimensional array that can capture the distribution of knowledge objects across text sentences. The rows of the grid correspond to the sentences in the text, whereas the columns correspond to the extracted knowledge objects from the text. For each object in the web page, the corresponding grid cell contains information about its grammatical role (S (Subject), O (Object), X (neither subject nor object) and gap, which signals the absence of the object in a given sentence.

The object grid is an extension of the popular entity grid representation for local coherence modelling proposed by Barzilay and Lapata (2008). The authors mainly extended the entity grid in three aspects to capture more information about the text. First, named entities and compound nouns are treated as head nouns in the entity grid. However, individual words cannot express a definite meaning and cannot describe the topic of the text explicitly. Thus, the object grid uses objects that are composed of terms and named entities instead of words. Second, named entities are divided into eleven classes (Person, Foundation, Project, etc.). These semantic entity types play important roles in distinguishing the category of the web resource. For example, news articles are likely to be about people and organisations. Third, the entity grid treats entities independently because it cannot capture the factor of lexical cohesion between entities. We address this problem by clustering entities semantically and by using the semantic chain to connect semantically related entities.

To construct the object grid, we need to deal with two things. First, we need to extract the objects and the relationships between them from text automatically. To extract the needed objects, we define a research ontology that organises various object types and relationships useful for the construction of the object grid. This ontology shows all types of named entities that we regarded as objects for object-based computation. The named entities mentioned in web pages are considered the object instances of certain types defined in the ontology. All object instances extracted from the text will be used for the object grid construction. Figure 1 shows part of the research ontology. To identify these objects and their relationships, we bring forward a series of methods, such as

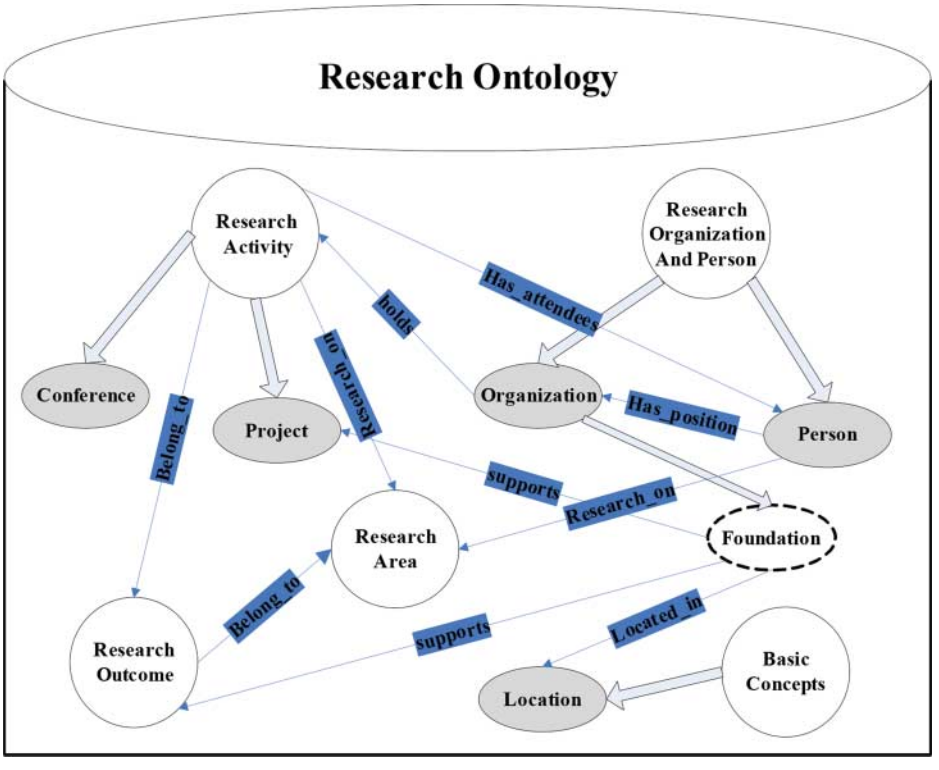

Figure 1. Structure of the research ontology.

tokeniser and sentence splitter, after basic nature language processes. First, the simplest method of object extraction is the dictionary-based approach. We collect some special object instances and some indication words that will pay an important role in rule construction. Limited by the dictionary size, the method is not flexible in the extraction of new objects. By contrast, some construction rules are based on object instances, such as parts of speech, syntax and structure. We filter the complete syntax phrases from a sentence for further analysis and design a rule-based model. To construct a rule, we have used two other modules that refer to the indication word identification and lexicon-syntactic pattern learning. For the objects that do not match any fixed pattern, we analyse the context feature and compute the similarity between the sentence containing the object and the sample according to the assumption that concepts that are semantically related tend to be near in terms of context as to that of plain text (Athanasios and Vangelis 2008; Zhang et al. 2009).

Second, we should preserve as much information of the objects as possible and construct the object grid. Basing on the lexicalised models for statistical parsing proposed by Collins (1997), we could mark the role of each extracted object in the text. During the annotation, we handled the co-reference of objects on the basis of the research of Ng and Cardie (2002). Subsequently, we constructed the object grid to represent the positional relation, syntactic relation and semantic relation among objects. Figures 2 and 3 show the example of object role annotation and object grid.

The method used in the current study in identifying the important objects in the web page is different from that used in other related research. In the present study, the important objects are divided into global objects and local objects, which are used to represent the topic and sub-topics of

1 [NASA's Voyager 1 spacecraft]$_S$ has entered a [new region]$_O$ between [our solar system]$_X$ and [interstellar space]$_X$.
2 [Data]$_S$ obtained from [Voyager]$_X$ over the last year reveal this [new region]$_S$ to be a kind of [cosmic purgatory]$_O$.
3 In [it]$_X$, [the wind of charged particles]$_S$ streaming out from [our sun]$_X$ has calmed, [our solar system's magnetic field]$_S$ has piled up, and [higher-energy particles]$_S$ from inside our [solar system]$_X$ appear to be leaking out into [interstellar space]$_X$.
4 "[Voyager]$_S$ tells us now that we're in [a stagnation region]$_X$ in [the outermost layer of the bubble]$_X$ around [our solar system]$_X$," said [Ed Stone]$_S$, [Voyager project scientist]$_X$ at [the California Institute of Technology]$_X$ in [Pasadena]$_X$.
5 "[Voyager]$_S$ is showing that what is outside is pushing back. We shouldn't have long to wait to find out what [the space]$_X$ between [stars]$_X$ is really like."
6 Although [Voyager 1]$_S$ is about [11 billion miles (18 billion kilometers)]$_O$ from [the sun]$_X$, it is not yet in [interstellar space]$_X$.
7 In the latest data, [the direction of the magnetic field lines]$_S$ has not changed, indicating [Voyager]$_S$ is still within [the heliosphere]$_O$, [the bubble of charged particles the sun]$_S$ blows around itself.

Figure 2. Example of object role annotation (Staff Writers 2011).

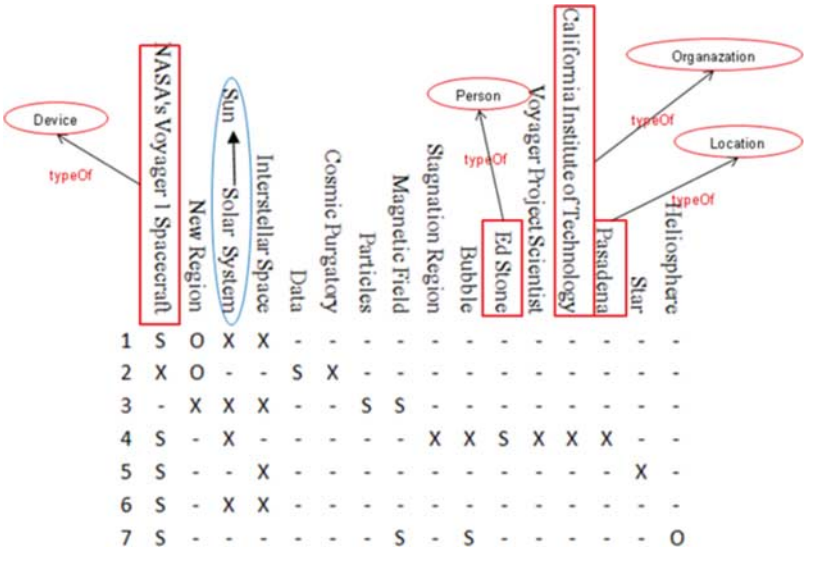

Figure 3. Object grid example.

the text, respectively. A fundamental assumption underlying our approach is that the distribution of entities in texts exhibits certain regularities reflected in grid topology. This assumption is not arbitrary. Some of these regularities have been recognised in centering theory (Grosz, Joshi and Weinstein 1995) and other entity-based theories of text. The global objects based on their distributed grid patterns, such as object cluster, object coherence, object density and object span, were identified. The anchor text and the 'meta' label of the web page are useful for this task. Given that the coherence is more visible inside the sub-topic than inside the topic, the text could be split into a number of semantic blocks by using this regularity. Therefore, local core objects could be identified based on their distributed patterns in each block.

3.2. *Object-based computation*

With the important objects identified, we could now implement computations, such as judging the intelligence value of web policy resources to support scientific policy-makers and categorising web pages into detailed classification.

To compute the value of web resources automatically, we should first understand the basic process of finding useful information implemented by human beings. Before the analysis, people usually retrieve many resources, including news. These resources are retrieved by using certain key words or by browsing certain parts of websites that have a close relationship with their topics. Humans will filter and sort these resources according to the resource type, source, title and abstract. If they find some words closely related to their target task of titles or abstracts, which are defined as intelligent sensitive words in this paper, they will decide to read these resources in depth to find more useful information. These sensitive words may refer to special persons, organisations, programmes, terms and so on. All of these sensitive words and their featured information are included in our object. Basing on the above processes, we have determined that some features could be used for automatic computation.

- Authority of source. If one piece of news is released in an official website, then this resource is more reliable than other resources from non-official websites.
- Content type of resource. If one resource is an extensive research report written by experts or research groups, then this report includes more valuable information than the regular news.
- Referred objects. If many important sensitive objects are mentioned in one resource, then this resource is considered more important than other resources.

However, most processes of judging the value of resources employ qualitative reasoning. To make our process more automatic and quantitative, we define five feature dimensions and propose corresponding computable indicators for important science and innovation policy resource identification. Table 1 shows the details of those five features and their corresponding indicators. Many indicators of these five features are related to important objects. Through the object grid and its featured information, such as semantic types and other attributions, the important objects identified will be used as items in vector computation. The weights of the feature vary and will be given by information experts.

The important objects are extracted to compute the indicators. Some computation methods are presented below.

(1) *Web authority*

The authority of web resources will be computed according to the five indicators in our method. Formula (1) shows the computation method.

$$\begin{aligned} Authority(D) = {} & Score(site) + Score(siteType) + Score(dir) + Score(dirType) \\ & + Score(country) \end{aligned} \tag{1}$$

Here, D represents the web page for computation, and $Authority(D)$ refers to its authority value. $Score(site)$ indicates the importance score of the website where the web page came from, and $Score(dir)$ is the score of the website directory where the web page was obtained from. $Score(siteType)$ and $Score(dirType)$ show the type of institute corresponding to the website and the directory type, respectively. The directory type is classified by the main resource type. $Score(country)$ is the score of the location of the institute. All scores used here are semi-automatically given through the works of information experts. The scores could be changed in terms f task.

(2) *Object relevancy*

We compute the object relevancy of one resource through four indicators, namely, object frequency ($F(O)$), object important score ($IS(O)$), object length ($L(O)$), and length of the main web content ($L(D)$). Different object instances with different semantic types are added together. As shown in Formula (2), D represents the web page for computation. Object *Relevancy* (D) is the object relevancy of the web page. $Fi(O)$ is the frequency that one object instance O occurs in the web page, and $Li(O)$ is the length of this object instance. Even though the features are similar, different objects have different importance scores. Thus, $ISi(O)$ indicates the importance level of object O. $L(D)$ refers to the length of the main content of a web page. In this formula, i is the number of

Table 1. Details of the five features and their corresponding indicators

Feature dimension	Corresponding indicators	Descriptions
Web source authority	Country type of source organisation. (Different country types have different levels of importance)	Examples include strong scientific country, BRICS[a], and developed country
	Type of source organisation. (Different organisation types have different levels of importance)	Examples include scientific management organisation, science foundation, research institute, and news site
	Importance of each organisation	Importance marks are given by intelligent analysers
	Type of source directory	Examples include strategy, research report, publication, news, and events
	Importance of each directory	Importance values are given by intelligent analysers
Content type of resource	Type of file	Refers to PDF, PPT, XLS, DOC, or HTML
	Length of the main web content	
	Ratio of main content and whole web	A higher ratio implies that the resource contains more information
	Feature words about content type in the title and full text	Some feature words could reflect the type of resources, such as 'annual report, budget, and research report'
	Type of source directory	Examples include strategy, research report, publication, highlight, news, press release, and events
Object (narrow)	Different objects and their importance	Objects here do not contain terms. Objects refer to people, organisations, conferences, programs, strategies, and so on. The importance marks are given by intelligent analysers first. Afterward, these marks are computed through the object grid
Science and innovation related terms	Domain-related terms	Core vocabularies of science and innovation. These terms could be used to compute the domain relation of resource.
	Domain hot terms	Hot topic of science and innovation. Resources talking about the hot topic may attract users
Policy related terms	Common words	These words are usually used with other types of terms or objects
	Scientific words	

[a]BRICS is the acronym for an association of five major emerging national economies: Brazil, Russia, India, China and South Africa.

important objects with different semantic types and values.

$$\text{Object Relevancy } (D) = \sum_{i=1}^{n} Fi(O) * Li(O) * ISi(O)/L(D).$$

Similar to object relevancy, science and innovation relevancy and policy relevancy can be computed through the corresponding types of the extracted word. To simplify computation, each indicator is computed separately and the results are normalised to the [0–1] zone.

We defined some rules, which contain several indicators from the same or different dimensions, with the help of information experts. These rules are separated into two groups, namely, important rules and unimportant rules. All rules and examples are related to the practical tasks of the intelligence analysis of NSL. For example, we offer more attention to the large scientific and technological powers to monitor the trend of scientific and technological policies. Therefore, the USA is more important than Nepal in this task. However, this preference varies with the task.

- Important:
- Important person + sci/tech innovation terms (e.g. Barack Obama + Sci&Tech|Innovation)
- important source + important country + Sci&Tech| Innovation terms (e.g. OECD + United States + Sci&Tech|Innovation)
- …
- Unimportant:
- important person + simple event news (e.g. Barack Obama + visiting …
- unimportant country + Sci&Tech| Innovation terms (e.g. Nepal + Sci&Tech|Innovation)
- …

Basing on the computed scores for each indicator and rule, we judge the value of science and innovation policy resources both quantitatively and qualitatively. If one web page acquires high scores in all the five features, then that web page will be very significant for the science and innovation policies. The computed intelligence value of a web page corresponds to the quantity of referred objects involved with the important or unimportant rules.

In addition to intelligence value judgment, we also employed important objects to conduct further resource classification, monitor conferences and research communities in a certain period of time, and so on. Basing on the important objects and their semantic type of information, we organise the web pages of the same objects into the same semantic type. We could further combine object frequency, position of objects in a document and document frequency on the basis of the extracted objects and their time feature. This combination requires a certain period of time to find the hot objects. More details on the computation processes are reported in another paper by Zhang et al. (2011).

4. Application

The authors implemented an RP system that monitors the websites of 86 science and innovation authority organisations, such as Office of Science and Technology Policy and Research Councils UK. The effectiveness of profiling the science and innovation policies of those institutes via the object-based computing method was demonstrated. All these organisations were chosen by science and innovation policy information experts. In this application, 11 information experts were involved. They are members of the scientific policy and strategy team of NSL in CAS. They chose 86 important organisations where they obtained resources for further intelligence analysis. Moreover, they provided the basic important objects with certain weight scores and rules that they have been recently concerned with most. With the help of the experts, we crawled related web resources in narrow scopes, constructed a basic gazetteer for object extraction and implemented qualitative judgment rules.

Using this system, we provided the newest important science and innovation policy web resources, automatically classified these resources, furnished important topics and objects in the most recent month, and kept track of the certain topics and objects within target organisations.

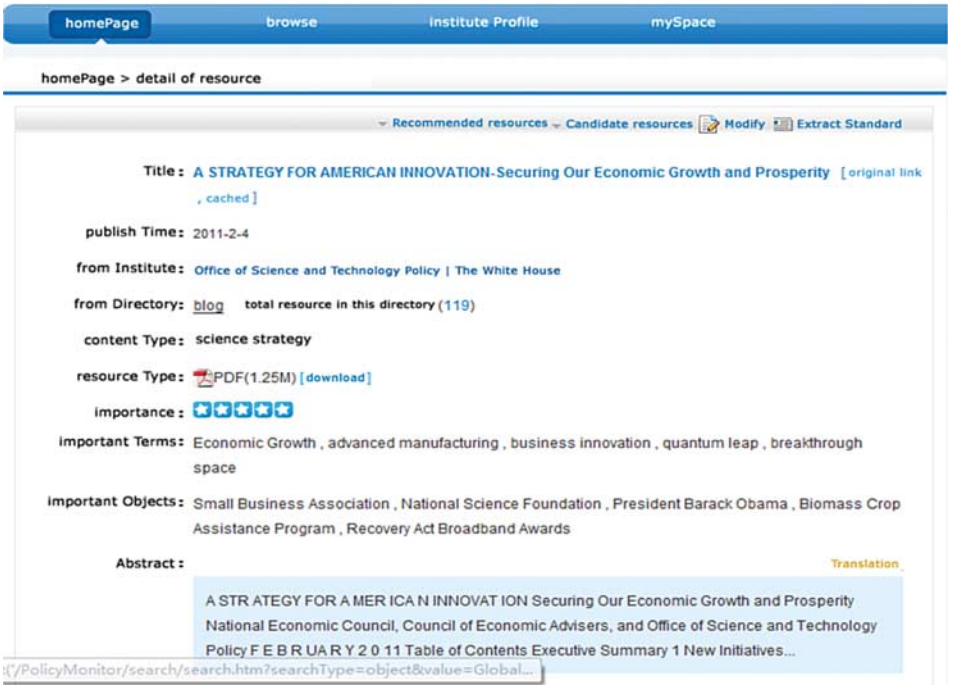

Figure 4. RP result of an important web resource.

Figure 5. Important object identification of an organisation.

According to the testing token by some intelligence experts, the method presents good performance. Figure 4 shows the RP result of one important web resource identified in our system. This web resource profiles its intelligence value, which is marked by stars, content type classification,

Table 2. Check result of recall and precision

Test amount of resource	Recall (%)	Precision (%)
100	65	27
300	89	50.2

key terms, key objects, related resources with same objects, and so on. Figure 5 shows the key terms and objects of the Office of Science and Technology Policy of the White House. Moreover, this figure presents the development trend of one object in different periods of time. More information could be obtained on the following URL: http://policyMonitor.las.ac.cn.

We invited some science and innovation policy intelligence experts to check the effectiveness of our object-based computing method. They assessed the precision and recall of the important resources for the science and innovation policy identified by our method. *Precision* refers to the ratio of the amount of important stars marked by our system and was identified by intelligence experts from the total test resource. *Recall* refers to the ratio of the amount of important resources identified by intelligence experts from the total test resource. We selected the top 100 important resources published from June 6, 2011, to June 16, 2011, and the top 300 important resources published from 1 July 2011 to 30 July 2011. The result is shown in Table 2.

The test result shows that our method could determine the most important resources for information experts. However, the performance of the exact importance level of each resource is not satisfactory. Moreover, more test resource situations show better performance in recall and precision.

5. Conclusions and future recommendations

This paper puts forward the object-based computation method for profiling science and innovation policies. According to our practice, the object-based computation method performed satisfactorily in identifying important resources.

Although we carried out some successful applications and received good evaluation results, much work needs to be done to widely apply our method in the future. Based on the test results, we will improve the method to determine a reasonable weight setting for each feature and adjust the algorithm for identifying the quality of important objects. Furthermore, we will determine a method of combining rules automatically on the basis of the existing important resources. All of these factors will affect the final mark and the amount of stars, which shows the importance level of the resources.

Acknowledgements

This paper is supported by the project titled 'Scientific Development Trend Detection System', funded by the Chinese Academy of Sciences (2009–2011) and 'The computing method of subject centrality of texts based on language network' supported by National Natural Science Foundation of China (Grant No. 61075047).

Notes on contributors

Zhixiong Zhang is a professor and the Assistant Director of National Science Library at the Chinese Academy of Sciences. He is also the Director of Information Systems, Library Department. He has published over 90 papers in journals,

conferences and workshops. His current interest areas include information extraction, web mining, research profiling, scientometrics, semantic web, and digital preservation.

Jianhua Liu is a PhD candidate of the National Science Library, Chinese Academy of Sciences. She specialises in text mining and scientometrics.
Yimin Zou is a lecturer at the College of Economics and Management, Zhejiang Normal University. He specialises in text mining and semantic web.

Jing Xie is a librarian of the National Science Library, Chinese Academy of Sciences. He specialises in distributed computing and semantic index.

Li Qian is a PhD candidate of the National Science Library, Chinese Academy of Sciences. He specialises in text mining and information visualisation.

References

Barzilay, R. and M. Lapata. 2008. Modeling local coherence: an entity-based approach. *Computational Linguistics* 34, no. 1: 1–34.

Choi, D.G., H. Le, and T. Sung. 2011. Research profiling for 'standardization and innovation'. *Scientometrics* 88, no. 1: 259–278.

Collins, M. 1997. Three generative, lexicalised models for statistical parsing. In *Proceedings of the ACL/EACL*, Madrid, Spain, July 1997, 16–23. Stroudsburg, PA: The Association for Computational Linguistics.

Gerd, S., H. Andreas, and B. Bettina. 2006. Semantic web mining: state of the art and future directions. *Web Semantics: Science, Services and Agents on the World Wide Web* 4, no. 2: 124–143.

Grosz, B.J., A.K. Joshi, and S. Weinstein. 1995. Centering: A framework for modeling the local coherence of discourse. *Computational Linguistics* 21, no. 2: 203–225.

Guo Y., L. Huang, and A.L. Porter. 2010. The research profiling method applied to nano-enhanced, thin-film solar cells. *R&D Management* 40, no. 2: 195–208.

Hicks, D., and G. Atlanta. 2004. *Global research competition affects US output*. School of Public Policy. Atlanta, GA: Georgia Institute of Technology.

Johanna, B. and S. Jan. 2007. Profiling academic research on digital games using text extraction tools. In *Proceedings of the Digital Games Research Association (DiGRA) Conference, 24–28 September 2007, Tokyo, Japan*, ed. Akira Baba, 714–729. Tokyo: DiGRA.

Johanna, B., R. Sami, S. Anne, and M. Petri. 2007. Enriching literature reviews with computer-assisted research extraction. Case: profiling group support systems research. Paper presented at the proceedings of the 40th annual Hawaii International Conference on System Sciences, February 2007, Hawaii.

Kleinberg, Jon. 2002. Bursty and hierarchical structure in streams. Paper presented at the proceedings of the eighth ACM SIGKDD International Conference on Knowledge Discovery And Data Mining (KDD'02), Edmonton, AB, Canada.

Lapata, Mirella and Regina Barzilay. 2005. Automatic evaluation of text coherence: models and representations. Paper presented at the proceedings of the 19th International Joint Conference on Artificial Intelligence, Edinburgh.

Ng, V. and C. Cardie. 2002. Improving machine learning approaches to coreference resolution. Paper presented at the proceedings of the ACL, July, Philadelphia.

Porter, A.L. and Nils C. Newman. 2011. Mining external R&D. *Technovation* 31, no. 4: 171–176.

Porter, A.L., A. Kongthon and J.C. Lu. 2002. Research profiling: Improving the literature review. *Scientometrics* 53, no. 3: 351–370.

President Obama. 2009. President Obama lays out strategy for American innovation. http://www.whitehouse.gov/the_press_office/president-obama-lays-out-strategy-for-american-innovation.

Sehgal, Aditya Kumar. 2007. Profiling topics on the web for knowledge discovery. PhD diss., University of Iowa.

Staff Writers. 2011. Voyager hits new region at solar system edge. *Space Daily*. http://www.space-travel.com/reports/Voyager_Hits_New_Region_at_Solar_System_Edge_999.html (accessed November 12, 2012).

Tegos, A., V. Karkaletsis, and A. Potamianos. 2008. Learning of semantic relations between ontology concepts using statistical techniques. In *Proceedings of the workshop on High-Level Information Extraction (HLIE 2008) at ECML-PKDD 2008, Antwerp, Belgium, 19 September 2008*. http://www-ai.cs.tu-dortmund.de/HLIE08/slides/03-tegos-HLIE_08_Tegos.pdf.

Zhang, Zhi-Xiong, Jian Xu, Jian-hua Liu, Qi Zhao, Na Hong, Si-zhu Wu, and Dai-qing Yang. 2009. Extraction knowledge objects in scientific web resource for research profiling. Proceeding of 2009 International Conference on Machine Learning and Cybernetics (ICMLC), July 12–15, Baoding.

Zhang, Zhixiong, Na Hong, Ying Ding, Jian-hua Liu, Jian Xu, and Dai-ling Yang. 2011. Research profiling based on semantic web mining. Paper presented at the International Council for Scientific and Technical Information Annual Conference (ICSTI 2011), Beijing.