Profiling Science & Innovation Policy by Object-based Computing

Zhang Zhixiong, Liu Jianhua, Zou Yimin, Xie Jing, Qian Li

National Science Library, Chinese Academy of Sciences, Beijing, China

Address: 33 Beisihuan Xilu, Zhongguancun, Beijing P.R. China, 100190.

Email: zhangzhh@mail.las.ac.cn

Zhang Zhixiong, PhD, Professor, Assistant Director of National Science Library, Chinese Academy of Sciences, and Director of Information System Department of the library. He is a standing committee member of IFLA (International Federation of Library Associations) IT Section, a member of Academic Committee for Digital Library Research and Development of the Library Society of China, and also as a professor of Graduate University of Chinese Academy of Sciences. He has published about 90 research papers on topics of information extraction, web mining, research profiling, information retrieval, ontology and semantic web, and digital preservation, etc.

Liu Jianhua, PhD. Candidate of National Science Library, Chinese Academy of Sciences, specialized in information extraction and text mining.

Zou Yimin, PhD. Candidate of National Science Library, Chinese Academy of Sciences, specialized in text mining and semantic web.

Xie Jing, M.S., librarian of National Science Library, Chinese Academy of Sciences, specialized in semantic web data management, distributed computing, and semantic index.

Qian Li, PhD. Candidate of National Science Library, Chinese Academy of Sciences, specialized in text mining and information visualization.
Profiling Science & Innovation Policy by Object-based Computing

Amount of named entities, which are regarded as research objects or objects, such as key initiatives & research programs, scientific strategies and policies, are embedded in many web pages from the websites of science & innovation institutes. These objects provide important information to extract intelligently from those pages. This paper brings forward an object-based computing method by using objects and their semantic information for profiling science & innovation policies. After extracting the objects and relations between them, we proposed an object grid to represent web pages which transferred the objects into machine-readable knowledge units with rich semantic information. Using the computational objects, we judged the intelligence value of web resources, and classified the policies into more detailed categories, such as strategic plan, R&D budget and so on. To test the effectiveness of profiling science & innovation policy by object-based computing method, this paper also conducted a research profiling experiment system.

Keywords: Research Profiling; Science & 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, which could be used to broadly scan the contextual literature information and depict the research context and research efforts wisely (Bragge, Sami, et al. 2007; Porter, Kongthon, et al. 2002). With the help of RP tools, one can easily identify related topics within the research domain, sketch out a “big picture” perspective on the research activity, understand the research community, get insight into how innovation is progressing, and map (graphically represent) topical interrelationships of a whole research area.

Much work about RP (Porter and Newman 2011; Choi, Lee, et al. 2011; Guo, Huang, et al. 2010; Bragge and Jan 2007; Hicks and Atlanta 2004) has been carried out. The data
they used for analysis are mainly on science, technology & innovation databases such as Web of Science and CSCD, etc. Few researchers use web resources for research profiling or adopt research profiling tools especially for web resources. Compared with the high-quality formal databases, the web resources are informal, unstructured, poor in semantic meaning, sometimes unreliable and unstable. However, the web resources still have advantages in terms of size, timeliness, accessibility and diversity. With the help of semantic web mining technologies (Stumme 2006), we could apply the profiling application into web resources. Aditya Kumar Sehgal (2007) proposed a profile-based approach to explore social networks of US senators generated from web data and compare with networks generated from voting data. But there is no in-depth study under the analysis of web contents.

National Science Library (NSL) is an information analysis services provider of Chinese Academy of Sciences (CAS). It is a very important task for the library to learn timely science & innovation policies and provide research reports to policy makers. Apart from the science, technology & innovation databases, some web pages released by key science & innovation related institutes, such as the Office of Science and Technology Policy of the White House, are also another important resources. Several projects have been carried out in NSL to support automatically or semi-automatically monitoring and profiling the science & innovation policies released in the websites of those key institutes. In this paper, the authors focus on web resources of some key institutes to implement science & innovation policy research profiling, then try to monitor and depict the development of science & innovation policies.

Based on science, technology & innovation databases, our project is similar to RP processes which require to deal with web pages obtaining, content extraction, analysis and visualization. Considering the great challenges to do RP upon 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 using the new approach. Section 5 gives concluding remarks and looks far ahead into the future work.

2. Main ideas of object-based computing

To present the main ideas of object-based computing, the meaning of object should be firstly explained. Through a lot of practices, the amount of named entities, such as scientific strategies and policies, key initiatives & research programs, key research institutes, which are embedded in the web pages, usually contained the core information of the certain web pages. It is very valuable for automatically mining intelligence from the web pages. For example, in a piece of news with the title of "President Obama Lays Out Strategy for American Innovation" (Obama 2009), there are two important named entities—— "President Obama" and "Strategy for American Innovation". These two entities both could present the main topics and value of the news directly. Hence, we consider these meaningful named entities mentioned above, as research object or object. Making use of these objects (meaningful named entities) and their inner relationships, we could further accomplish the certain computation for knowledge discovery. Therefore, an object-based computing method is used to identify the objects and the relationships between them, which are extracted from web pages, so as to carry out computation for profiling science & innovation policies. In this method, there are four main ideas:
• Crawl web pages about science & innovation policy from the websites of key institutes continuously, such as the Office of Science and Technology Policy and U.S National Science Foundation, with the purpose of monitoring the changes of science communities. It is important to choose the institutes correctly in view of the authority and reliability of web resources obtained. The experts in the fields will be involved in institute filter process.

• Turn the crawled free texts into time-stamped objects (when objects occur, that is objects with temporal feature) through information extraction and identify important objects to support object-based computing. After related web pages crawled from targeted institutes’ website, we need transfer the unstructured free texts into the structured and machine-readable object collections. Because of the dynamic of web resources, we add the temporal feature of objects when we label these extracted objects. Take “July 13, 2010, White House Announces National HIV/AIDS Strategy” for example, we turn it 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 still record other context information of the objects, such as source, syntax, and position in the text. With the information of objects extracted from text, we construct an object grid to identify the important objects, which will be explained in details in the following section.

• Build a large scale knowledge base based on time-stamped objects to achieve semantic mining of the related topics. The large scale knowledge base with time-stamped objects is important for the object-based computing. In our method, all
the extracted objects and their semantic information are preserved in relation databases. These objects and their relationships representation model can be extended to content analysis tasks involving a temporal dimension, such as burst topic detection (Kleinberg 2002).

- Profiling the status of science communities by using information visualization. Based on enormous objects, we identify new science & 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, R&D budget and organization restructure), outline the hot topics in a period of time, depict current active activities and players in the institutes, cluster the related policies of the institutes.

As web crawling and information visualization are not the key issues of object-based computing, this paper specifically focus on object identification and object-based computing.

3. Key issues
As it mentioned above, object identification, object-based computing are the key issues of this paper. Firstly, we discuss the important objects identification.

3.1. Important objects identification
As it is posed in the "main ideas of object-based computing" section, amount of named entities embedded in web pages usually carry core information of certain web pages. They could present the main topics and value of the news directly. It is very important for us to extract them from the unstructured web pages. For this purpose, the concept of object grid, which transfers the objects embedded in the web into structured and
machine-readable knowledge units, will be involved in this paper. Object grid is a two-dimensional array which can capture the distribution of knowledge objects across text sentences. The rows of the grid correspond to the sentences in the text, while the columns correspond to 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), -( gap which signals the object’s absence from a given sentence)) in the given sentence.

Object grid is an extension of the popular entity grid representation for local coherence modelling which is proposed by Barzilay and Lapata (Barzilay, Lapata, 2008). The authors mainly extend the entity grid in three aspects which can capture more information about the text. First, named entities and compound nouns are treated as their head nouns in the entity grid, but individual words cannot express a definite meaning and describe the topic of the text explicitly. So, Object grid uses objects which are composed of terms and named entities instead of words. Second, named entities are divided into eleven classes (Person, Foundation, Project, et al.), 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 organizations. Third, the entity grid treats entities independently, as it cannot capture the factor of lexical cohesion between entities. We address this problem by clustering entities semantically and using the semantic chain to connect semantically related entities.

To construct the object grid, we need to deal with two things. Firstly, we need extract the objects and the relationships between them from text automatically. To extract the needed objects, we define a research ontology which organizes various object types and relationship useful for object grid construction. This ontology shows all types of named entities that we regarded as objects for object-based computing. The named entities
mentioned in web pages are considered as the object instances of certain types defined in the ontology. All the object instances extracted from 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 series methods after basic nature language process such as tokeniser and sentence splitter. Firstly, the simplest method of object extraction is dictionary-based approach. We collect some special object instances and some indication words which will pay an important role in rule construction. Limited to the size of dictionary, it is not flexible in new objects extraction. On the other hand, there are some construction rules of object instances, such as part-of-speech, syntactic, structure and so on. We filter the complete syntax phrases from a sentence for further analysis and design a Rule-based model. To construct rule, we used other two modules which refer to indication word identification and lexicon-syntactic pattern learning. For the objects that don't match any fixed pattern, we analyzes the context feature and compute similarity between the sentence containing object and the sample one, according to the assumption that concepts which are semantically related, tend to be near as context in a plain text (Athanasios and Vangelis 2008; Zhang, Xu, et al. 2009).

Figure 1 Part structure of the research ontology

Secondly, we should preserve as much information of objects as possible and construct the object grid. Based 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 still deal with the co-reference of objects based of Ng and Cardie's (Ng and Cardie, 2002) research. Then the authors construct the object grid to represent the positional relation, syntactic relation and semantic relation among the objects. Figure 2 and figure 3 show the example of object role annotation and object grid.

Figure 2. Example of object role annotation (Staff Writers of Space Daily 2011)
Figure 3. Object grid example The method which identifies the important objects in the web page is different with other related researches. The important objects are divided into the global objects and local objects which are used to represent the topic and sub-topics of 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 recognized in Centering Theory (Grosz et al., 1995) and other entity-based theories of text. The authors identify the global objects based on their distributed patterns in grid, such as the features of objects cluster, objects coherence, objects density and objects span, etc. In addition, anchor text and the “meta” label of the web page are also useful for this task. As we know that the coherence is more visible inside the sub-topic, the text could be split into a number of semantic blocks using this regularity. So, local core objects could be identified based on their distributed patterns in each block.

3.2. Object-based computing

After the important object identified, we could apply them to implement some computation, such as judge the intelligence value of web policy resources to support scientific policy-makers, and classify the web pages into detailed classification.

To compute the value of web resources automatically, we should understand the basic process of finding useful information implemented by human beings firstly. Before the analysis, people usually retrieve a lot of resources, including news, by using certain key words or browsing certain parts in the websites which have close relationship to their topics. Then they 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 intelligence sensitive words in this paper, they
will decide to read these resources in depth to find more useful information. Such sensitive words refer to some special persons, organizations, programs, terms and so on. All of these sensitive words and their feature information are included in our object. Through above processes, we believe that there are some features we could use for the automatic computing.

- The authority of source. If one piece of news is released in an official website, it is more reliable than other resources from non-official websites.

- The content type of resource. If one resource is a deep research report written by many experts or research groups, it includes more valuable information than the regular news.

- The referred objects. If there are many important sensitive objects mentioned in one resource, it is considered to be more important than other resources.

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

Table 1. Details of five features and their corresponding indicators

To compute the indicators above, the previous work of extracting the important objects
will be of great importance. Some computation methods will be presented below.

(1) Web authority

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

\[ \text{Authority(D)} = \text{Score(site)} + \text{Score(siteType)} + \text{Score(dir)} + \text{Score(dirType)} + \text{Score(country)} \]

In this formula, \( D \) is one web page for computing and \( \text{Authority (D)} \) refers to its authority value. \( \text{Score (site)} \) presents the important score of the website where the web page came from while \( \text{Score (dir)} \) means the score of directory in the website where the web page was from. \( \text{Score (siteType)} \) and \( \text{score (dirtype)} \) show the type of institute corresponding to the website and the type of the directory separately. The type of the directory is classified by the main resource type. \( \text{Score (country)} \) is the score of the location of institute. All the scores used here are semi-automatically given. Information experts will be involved in this work. The scores could be changed in term of task.

(2) Object relevancy

We compute the object relevancy of one resources through four indicators, including object frequency \( \text{\( F(O) \)} \), object important score \( \text{\( IS(O) \)} \), object length \( \text{\( L(O) \)} \), and Length of the main content of web\( \text{\( L(D) \)} \). Different object instance with different semantic type will be added together. As shown in Formula (2) below, \( D \) means one web page for computing. Object Relevancy \( \text{(D)} \) is the object relevancy of the web page. \( \text{\( F(O) \)} \) is the frequency that one object instance \( O \) occur in the web page, while \( \text{\( L(O) \)} \) is the length of this object instance. Be similar to features, different object has different important score. So the \( ISi(O) \) indicates the important level of object \( O \). At last, \( L(D) \) refers to the length of main content of a web page. In this formula, \( i \) is the number of the important objects with different semantic types and value.
ObjectRelevancy (D) = \sum_{i=1}^{n} F_i(O) * L_i(O) * IS_i(O) / L(D)

Be similar to the object relevancy, the science & innovation relevancy and policy relevancy could be computed through corresponding types of the extracted word. To simplify the computation, each indicator is computed separately and makes the results normalizing to [0-1] zone.

Besides, we define some rules, which contain several indicators from same or different dimensional, with the help of information experts. These rules are separated into two groups—important rules and unimportant rules. All the rules and examples are related to the practical tasks of intelligence analysis of NSL. For example, we pay more attention to the big scientific and technological powers to monitor the trend of scientific and technological policies. Hence, we believe that America is more important than Nepal in this task. It could vary from different tasks.

• 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 + America +Sci&Tech|Innovation)
  • ...

• un-important:
  • important person+simple event news (e.g. Barack Obama + visiting ......
Based on these computed scores of each indicator and rules, we judge the value of science & innovation policy resources both quantitatively and qualitatively. If one web page get higher score in all the five features, it will be very important for the science & innovation policies. If one web page contains more referred objects involved in important rules or un-important ones, it will obtain the corresponding intelligence value. Besides intelligence value judgment, we also make use of the important objects to conduct further resources classification, monitor the conferences or research communities in a certain period of time and so on. Based on the important objects and their semantic type of information, we organize the web pages of the same objects into the same semantic type. Based on the extracted objects and their time feature, we could further combine object frequency , position of object in document. and document frequency which occur in a certain period time to find the hot objects More details about the computation processes are written in another paper of Zhang Zhixiong (Zhang, Na, et al. 2011).

4. Application

As an application practice, the authors implement a research profiling system which monitors websites of 86 science & innovation authority organizations, including Office of Science and Technology Policy, Research Councils UK and etc., to show the effectiveness of profiling science & innovation policy in those institutes used by object-based computing method. All these organizations are chosen by science & innovation policy information experts. In this application, 11 information experts are involved.
They are members of scientific policy and strategy team of NSL in CAS. According to their research task, they choose 86 important organizations where they obtain resources for further intelligence analysis. Besides, they provide us the basic important objects with a certain weight score and rules they concerned most recently. With the help of experts, we crawl related web resources in narrow scopes, construct basic gazetteer for object extraction, and carry out qualitative judgement rules.

In this system, we provide the newest important science & innovation policy web resources, automatically classified these resources, important topics and objects in recent one month, and keep track of the certain topics and objects within target organizations and so on. According to the testing token by some intelligence experts, the method presents a good performance. Figure 4 shows the RP result of one important web resource identified in our system. It profiles its intelligence value which is marked by stars, content type classification, 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, it presents the development trend of one object in different period of time. More information could be obtained on the following URL: http://policyMonitor.las.ac.cn.

Figure 4. RP result of one important web resource

Figure 5. Important object identification of organization

To check the effective of our object-based computing method, we invited some science & innovation policy intelligence experts to assess precision and recall of the important resources about science & innovation policy identified by our method. The precision refers to the ratio of amount of important stars marked by our system identified by intelligence experts to total test resources. The recall refers to the ratio of amount of important resources identified by intelligence experts to total test resources. We selected
top 100 important resources published during June 6th, 2011 to June 16th, 2011 and top 300 important resources published during July 1st, 2011 to July 30th, 2011. The result is showed in table 2.

Table 2. check result of recall and precise

The test result shows that our method could find out the most important resources for information experts. But the performance of the exact important level of each resource is not good enough. Besides, more test resources situation shows better performance both of recall and precision.

5. Conclusions and future

This paper put forward the method of object-based computing for profiling science & innovation policy. According to our practice, our object-based computing method reveals a good performance in identifying the important resources.

Although we carried out some successful application and received a good evaluation results, there is still a lot of work to do for its more efficient use in the future. According to the test results, we will improve the method for reasonable weight setting of each feature and adjust the algorithm for identifying the quality of important object. Furthermore, we will learn more combining rules automatically based on the existing important resources. All of these factors will affect the final mark and the amount of stars which shows the important level of the resources.

Acknowledgements

This paper is supported by a project, which is titled "Scientific Development Trend Detection System", supported by Chinese Academy of Sciences (2009-2011).

References
Aditya Kumar Sehgal. 2007. Profiling topics on the Web for knowledge discovery. dissertation, University of Iowa


Jon Kleinberg. 2002. Bursty and hierarchical structure in streams. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD’02)


Ng, and C. Cardie. 2002. Improving machine learning approaches to coreference resolution. Proceedings of the ACL.

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


<table>
<thead>
<tr>
<th>Feature dimension</th>
<th>corresponding indicators</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Web source authority</strong></td>
<td>country type of source organization, (Different country type has different importance)</td>
<td>Such as strong scientific country, BRICS, developed country.</td>
</tr>
<tr>
<td></td>
<td>type of source organization (Different organization type has different importance)</td>
<td>Such as scientific management organization, science foundation, research institute, news site.</td>
</tr>
<tr>
<td></td>
<td>Importance of each organization</td>
<td>The importance marks are given by intelligent analyzers.</td>
</tr>
<tr>
<td></td>
<td>type of source directory</td>
<td>Such as Strategy, Research report, Publication, news, events...</td>
</tr>
<tr>
<td></td>
<td>Importance of each directory</td>
<td>The importance marks are given by intelligent analyzers.</td>
</tr>
<tr>
<td></td>
<td>type of file</td>
<td>Refer to PDF, PPT, XLS, DOC, HTML</td>
</tr>
<tr>
<td></td>
<td>Length of the main content of web</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Ratio of main content and whole web</td>
<td>Higher the ratio, the resource contains more information.</td>
</tr>
<tr>
<td></td>
<td>Feature words about content type in title and full text</td>
<td>Some feature words could reflect the type of resources such as &quot;annual report, budget, research report&quot;.</td>
</tr>
<tr>
<td></td>
<td>type of source directory</td>
<td>Such as Strategy, Research report, Publication, highlight, news, press release, events...</td>
</tr>
<tr>
<td><strong>Object (narrow)</strong></td>
<td>different objects and their Importance.</td>
<td>Object here doesn’t contain terms. It refers to person, organization, conference, program, strategy and so on. The importance marks are given by intelligent analyzers firstly, then they are computed through object grid.</td>
</tr>
<tr>
<td><strong>science and innovation related terms</strong></td>
<td>Domain related terms</td>
<td>Core vocabularies of science and innovation. These terms could be used for computing the domain relation of resource.</td>
</tr>
<tr>
<td></td>
<td>Domain hot terms</td>
<td>Hot topic of science and innovation. Resources talking about hot topic may attract users.</td>
</tr>
<tr>
<td><strong>policy related terms</strong></td>
<td>Common words</td>
<td>These words are usually used with other type terms or objects.</td>
</tr>
<tr>
<td></td>
<td>Scientific words</td>
<td></td>
</tr>
</tbody>
</table>
Table 2. Check result of recall and precise

<table>
<thead>
<tr>
<th>Test amount of resource</th>
<th>Recall</th>
<th>Precise</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>65%</td>
<td>27%</td>
</tr>
<tr>
<td>300</td>
<td>89%</td>
<td>50.2%</td>
</tr>
</tbody>
</table>

Figure 1 Part structure of the research ontology

Figure 2. Example of object role annotation (Staff Writers of Space Daily 2011)
Figure 3. Object grid example

Figure 4. RP result of one important web resource
Figure 5. Important object identification of Organization