Identify the Intelligence Values of Web Resources Based on Knowledge Objects Grid

Zhang Zhixiong, Zou Yimin, Liu Jianhua, Xie Jing, Qian Li
National Science Library, Chinese Academy of Sciences, Beijing,China
Address: 33 Beisihuan Xilu, Zhongguancun , Beijing P.R.China ,100190.
Email: zhangzx@mail.las.ac.cn

Abstract:

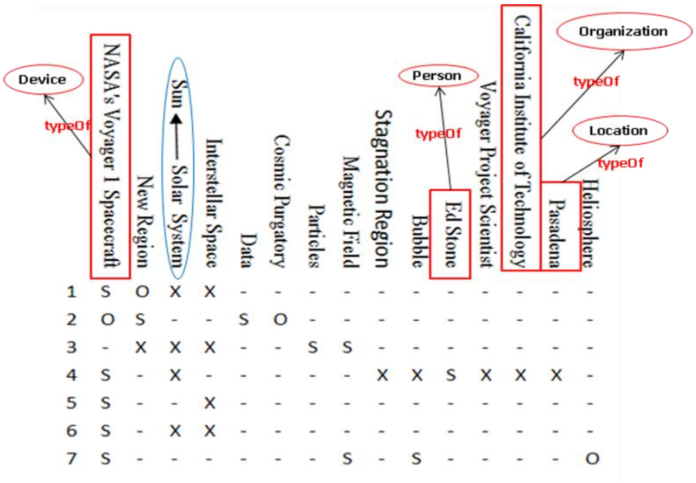
Several projects have been carried out in the Chinese Academy of Sciences to support automatically (semi-automatically) monitoring and profiling the S&T research activities based on web resources，which are published by some key institutes (such as some national administrative offices, research councils, funding agencies, and leading research institutes). It is very important to identify the intelligence value of those web resources to help information analyst to select valuable resources from a large number of gathered web resources.

In this paper, the authors bring forth a new method for judging the intelligence value of web resources based on a knowledge object grid. Knowledge objects are terms and several kinds of named entities embedded in web pages, such as the names of science strategies, research programs, research institutes, etc. Knowledge 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 text, while the columns correspond to extracted knowledge objects from text. For each knowledge object, the corresponding grid cell contains information about its grammatical role and other relations in the given sentence.

Knowledge object grid (Figure 1) is an extension of the popular entity grid representation for local coherence modeling which is proposed by Barzilay and Lapata. The authors of this paper mainly extend the entity grid in three aspects which can capture more information about the text. Firstly, head nouns of named entities and terms take place of the full named entities and terms in entity grid. In our opinion, single words can't present definite meanings and topics of one text explicitly. Therefore, the authors use knowledge objects, which consist of full terms and named entities, to replace single words to form knowledge object grid. Secondly, the authors defined eleven semantic classes for named entities, such as Person, Foundation and Project. These semantic classes will play important roles in distinguishing the category of web resource. For example, news articles are likely to be about people and organizations. Thirdly, the entity grid treats entities independently which could not capture the lexical cohesion between entities. We address this problem by clustering entities semantically and using semantic chain to connect related entities.

To identify the intelligence value of a web page, the authors need construct the knowledge object grid and identify the core knowledge objects in the web page. First of all, knowledge objects should be extracted from the web page. During the object extraction process, the authors parse the syntactic roles of each object, i.e. S(Subject),

O(Object), X(Other), -(NULL)), in each sentence automatically. To solve the co-reference of objects and identify related entities in semantic, WordNet and S&T Ontology which is built in the authors' other project are involved. Finally, the authors construct the knowledge object grid to represent the positional, syntactic and semantic relations among the objects.



	NASA's Voyager 1 Spacecraft	New Region	Sun ← Solar System	Interstellar Space	Data	Cosmic Purgatory	Particles	Magnetic Field	Stagnation Region	Bubble	Ed Stone	Voyager Project Scientist	California Institute of Technology	Pasadena	Heliosphere
1	S	O	X	X	-	-	-	-	-	-	-	-	-	-	-
2	O	S	-	-	S	O	-	-	-	-	-	-	-	-	-
3	-	X	X	X	-	-	S	S	-	-	-	-	-	-	-
4	S	-	X	-	-	-	-	-	X	X	S	X	X	X	-
5	S	-	-	X	-	-	-	-	-	-	-	-	-	-	-
6	S	-	X	X	-	-	-	-	-	-	-	-	-	-	-
7	S	-	-	-	-	-	-	S	-	S	-	-	-	-	O

Figure 1 A fragment of knowledge object grid .

Note: an example of knowledge object grid based on the text of "Voyager Hits New Region at Solar System Edge" from SpaceDaily.

The method which identifies the core knowledge objects in web page is different from other related researches. The core knowledge 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 regularities have been recognized in Centering Theory, Zipf's Law 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, coherence, density and span. In addition, anchor texts and the "meta" labels of web page are also useful for this task. As we know, the coherence is more significant inside the sub-topic. So, text could be split into a number of semantic blocks using this rule; and then local core knowledge objects could be identified based on their distributed patterns in each block.

Based on the identified core knowledge objects, the authors compare the core knowledge objects in each web resource and calculate the importance of web resource that the objects occur. Combined with other information from the web resources, such as the format, source, domain relevancy, and user interests profiles, the author implement an experiment system to identify the intelligence value of web resource.

Keywords: Research Profiling; Intelligence Value; Knowledge Object Grid；Knowledge Extraction.