Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009

EXTRACTION KNOWLEDGE OBJECTS IN SCIENTIFIC WEB RESOURCE FOR RESEARCH PROFILING

ZHI-XIONG ZHANG[1], JIAN XU[1], JIAN-HUA LIU[1], QI ZHAO[1], NA HONG[1], SI-ZHU WU[1], DAI-QING YANG[1]

[1]National Science Library, Chinese Academy of Sciences, Beijing100190, China
Email: Zhangzhx@mail.las.ac.cn, xujian@mail.las.ac.cn, liujh@mail.las.ac.cn, zhaoqi@mail.las.ac.cn, hongn@mail.las.ac.cn, wusizhu@mail.las.ac.cn, yangdaiqing@mail.las.ac.cn

Abstract:

Research profiling is a large-scale analysis method based on the literature information to depict the state-of-the-art scientific researches wisely. Much research profiling work has been carried out based on formal, structured scientific publications (such as Chemical Abstracts) while there is much information contained in abundant, unstructured scientific web resources. In order to profiling research based on those unstructured scientific web resources, the authors bring forth a solution that tries to extract useful knowledge objects from them. Three kinds of knowledge objects have been extracted: (1) research related objects (research objects), such as institutes, scientists, projects, conferences etc.; (2) the relationships between the research objects which reflected in scientific web resources, such as one scientist worked in one institute; (3) the terms which indicate the topic of the research areas. The authors implemented the knowledge objects extraction system and did some experiments to test and evaluate the effect of this system.

Keywords:

Knowledge Objects; Research object Extraction; Relation Extraction; Research Term Extraction; Research Profiling

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 to depict research context and research efforts wisely [1]. In much work about RP that has been carried out, such as [2][3][4], raw data used contain projects (e.g., NSF projects), citations, patents, publications (e.g., Chemical Abstracts) and other structured data based on formal scientific publications such as periodical papers, conference proceedings and patents. However, in real world, there is still much information contained in informal web resources. Those informal ones are open, dynamic, rich, cross-domain and could reflect the development of scientific researches immediately. On the other hand, they are unstructured, hard to define boundary, unreliable，unstable. All features are challenges to RP tasks based on web resources.

Nowadays, the authors of this paper are undertaking a project named Science Monitoring and Evaluation based on scientific web resources, which is funded by National Key Technology R&D Program in the 11th Five Year Plan of China. The project's goal is to (semi) automatically monitor and evaluate the development trends of important scientific research institutions. RP based on scientific web resources is the heart of our tasks. So, the most important problem we faced is how to extract structured semantic knowledge objects from unstructured text. The paper will show the complete methodology of knowledge objects extraction from scientific web resources for RP.

The rest parts of paper are organized as follows. Section 2 presents the knowledge objects need extracting in scientific web resources according to RP goal. Section 3 and section 4 describes the methodology of knowledge object extraction. Section 5 gives conclusions of this paper.

2. Knowledge objects in scientific web resources

For RP tasks, we defined a research ontology which organizes all research object and relation classes together. After research object and relation instances extracted, they can be attached to classes defined in ontology and be used for knowledge repository construction, which will benefit reasoning based on objects and relations extracted.

The main classes in ontology include *Research Activity*, *Research Outcome*, *Research Organization and Person*, *Research Establishment* and *Basic Concepts*. These classes can be classified further into more specific concepts. For example, *Research Activity* includes *Project, Conference, Lecture, Research Award, Experiment, Investigate* and *Train*. In addition to object classes, it also describes research object attributes and relations such as

978-1-4244-3703-0/09/$25.00 ©2009 IEEE

has_attendees relation between *Research Activity* and *Person*, *supports* relation between *Foundation* and *Project*, etc. Part of the research ontology is presented in figure 1. In this figure, circles denote top classes while ellipses with grey shadow denote classes belonging to top one. The other ellipses with broken line denote classes belonging to grey ones. Besides, thin arrows denote relations between classes and thick ones denote class hierarchy relations.

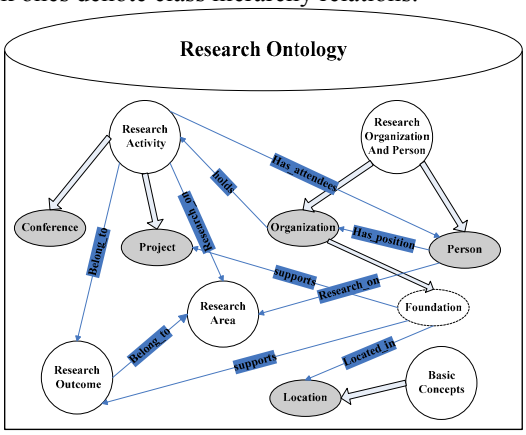

Figure 1. Structure of the research ontology

According to the ontology, knowledge objects which should be extracted include two aspects. Firstly, research can be seen as the interactive course among related research activities, outcomes, organizations and persons. Structured instances of such object and relation classes are the base of further analysis. Therefore, research objects and relations are one important aspect of knowledge object for research description. Secondly, term represents a kind of communication manner among scientists. It is a very useful semantic unit because it can represent single domain concept such as entity, process or function independently and show high relativity in certain theme [5]. Several terms can describe main idea of a document briefly. For further analysis on distributed structure of domain themes, research term is another important aspect.

3. Method of Research objects and their relations extraction

Recently, research in knowledge extraction has made some developments. This part presents our method of research object and relation extraction in scientific web resources based on existing systems and methods.

3.1. Research objects and their relations extraction

It has made great success in named entity extraction and formed general entity extraction framework which includes Person, Location, Organization, Data and Numerical [6]. But, the objects need to be extracted in our project are different from the general ones.

Firstly, many research objects are short of distinct indication. Take "*University of Sheffield*" for example, it can be extracted as a *university* easily since it consists of capital letter and indication word "*university"*. But there are few indications in the context of journal objects such as "*Investigative Ophthalmology and Vision Science*". It is difficult for some traditional extraction approaches such as rule based approach and context-based statistical approach to extraction them through their indication words.

Secondly, they have more complicated format than general ones. There are usually many words with different part-of-speech in one instance. For example, the conference "*European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*" consists of more than 6 words which include preposition, noun, gerund and other conjunction. All of these make it difficult to judge semantic boundary.

Thirdly, Research objects often characterized with variability and ambiguity. Along with official and standard names, a range of variants, synonyms and alternative names are used to refer to the same objects. For example, *MIT* and *Massachusetts Institute of Technology* are all used to denote same university object. Besides, ambiguity is another challenge for research object extraction. Some names may have multiple independent meanings. Take "*ACE*" for example, it may be a conference object "*Automatic Content Extraction*" or controlled language "*Attempto Controlled English*". All of these cause disambiguation difficulty.

According to the differences between research objects and general ones, we use various integrated approach such as lexical-pattern approach and statistical approach for research objects extraction. Besides, we choose some matured open source software such as GATE (General Architecture for Text Engineering) [7] and Stanford Parser [8] to provide basic NLP support. The whole technical framework is presented in figure 2.

In this figure, the blue modules are implemented using GATE's plug-ins directly which contain *Tokenize*, *Sentence splitter* and *part-of-speech*. These modules will provide basic NLP information of words such as kind, orthographic, etc. The red ones are developed based on GATE's resources or designed by ourselves while the green one denotes the out referring database.

The detail of each module is presented as follows:

(1) *Gazetteer Reorganization Module*. The simplest method of object extraction is dictionary-based approach. Limited to the size of dictionary, it is not flexible in new objects extraction. So we just use it for some special object extraction here and list some indication words in dictionary which will pay an important role in rule construction.

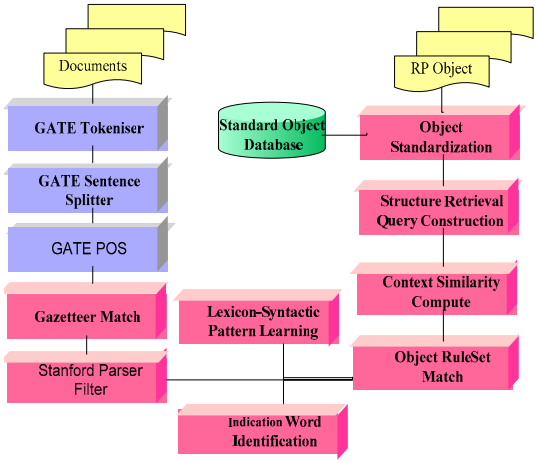

Figure 2. Framework of research object extraction

(2) *Stanford Parser Filter module*. Usually, research objects are presented by noun phrases. So we designed this module to filter the complete syntax phrases from a sentence for further analysis.

(3) *Object Rule Set module*. Rule-based approach is another important one in object extraction. In the rules, all kind of information of words such as part-of-speech, kind and so on will play important roles. To construct rule, we used other two modules which refer to indication word identification and lexicon-syntactic pattern learning.

(4) *Context Similarity Compute*. Besides dictionary-based and lexical-pattern approaches, some objects don't match any fixed pattern. According to the assumption that concepts which are semantically related, tend to be near as context in a plain text [9], we designed this module which contains two parts. One analyzes the context words' feature and the other compute similarity between the sentence containing research object and the sample one.

(5) *Hierarchical Structure Retrieval Query Construction*. According to Hearst model [10], the instance and concept often have some fixed expressions. For example, "*A is a kind of B*". Thanks to such definition, we constructed some hierarchical structure retrieval query to find the correct type of some research objects based on the feedback of search engine.

(6) *Object Standardization*. Based on analysis of the extraction result, we constructed a standard object map database for Object Standardization which contains the general and standard expression pairs.

3.2 Method of relation extraction

Though there are different views of relation extraction, they have same tasks which include relation element and relation mark identification, semantic type of relation judgment. Relation element refers to the research objects extracted from text while relation mark connects research objects. Relation mark may be just a kind of syntax form (such as possessive) or certain word and phrase (such as preposition and verb).

Due to the syntax features, different methods are chosen for relation extraction. We propose relation triple construction based on pattern and non-pattern methods and all relation triples constructed here are limited in the same sentence for reducing extraction difficulty.

To some relation triples, their elements and relation marks are fixed relatively such as the position in sentence and semantic information. For example, in "*The IBM team's*" and "*Nathan Myhrvold，Microsoft's chief scientist*", the employ relation is presented through possessive. So we could expand some relation rule set based on Hearst pattern. Take employ relation between organization and person for example. It usually has some fixed expressions as follows:

"<research person>, <position> of <organization>"
"<position>< research person >of<organization>"
"< research person> (<organization>'s <position>)"
"<organization>'s <position>< research person>"

We could extract the employ relation triples between organization and research person through *patterns method.*

In addition to some relations that have fixed rules, there are more have no rules because of various expressions of relation mark and flexible position of relation element and mark. So we need the *non-pattern Relation Triple Construction method* to deal with them.

Based on analysis of a lot of examples, it is found that the syntax function of relation element or mark is fixed relatively. Besides, there are some rely-link between relation element which refers to the distance between relation elements and mark. So we design an algorithm named *RelaPair* to construct the relation triples (Figure3). The whole flow is presented as follows.

(1) Input the document after tokenize, sentence splitting and research object extraction. Deal with each spitted sentence and construct two empty relation triple lists for preposition mark and verb mark.

(2) Judge if the sentence contains more than one research object. If not, give up this sentence and return to step (1) for next sentence. Skip to step (3) if containing one while to step (4) if contains more than two objects.

(3) Judge if there is preposition in research object. If not, give up it and go on next sentence. Else, judge if the noun phrases in the two side of preposition have semantic information (means the nesting semantic annotation). If not, return to step (1) for next sentence. Else, construct the relation triple and add it to preposition relation triple list.

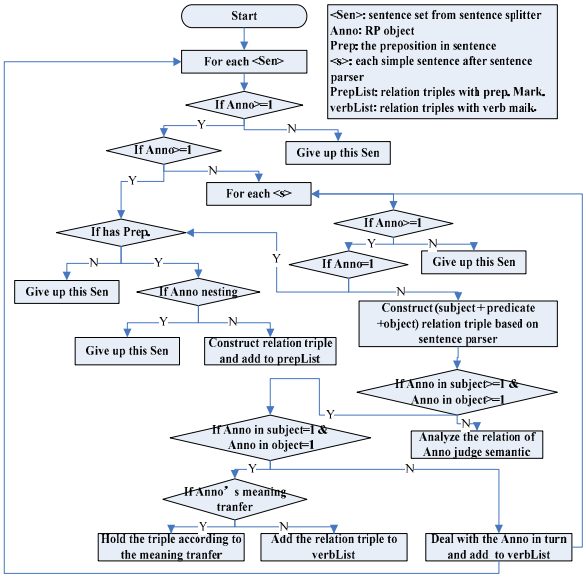

Figure 3. Flow of RelaPair algorithm

(4) Obtain the simple sentence (it refers to the sentence contains subject, predicate and object without any clause) through Parser. We could get the second circle deal point.

(5) Judge if the simple sentence contains more than one research object. If not, give up this one and return to analyze next one. Skip to step (3) if containing one while to step (6) if contains more than two objects.

(6) Obtain the (subject, predicate, object) relation and construct relation triples. Go to step (7)

(7) Analyze the (subject, predicate, object) relation triple. If there is at least one research object existing both in subject and object phrases, skip to step (8) while skip to step (9) if all research objects exist in the same phrase.

(8) If there is only one research object in both subject and object phrases, judge if their meaning is transferred such as possessive. If not, construct relevant relation triples and add them into verb relation triple list. If meaning transfer exists, decide to give it up or not according to the transfer degree. If there is more than one research object in both phrases, deal them in turns and pay attention to the meaning transfer caused by apposing.

(9) Analyze the research objects in phrases and judge the semantic type of relation triples through semantic information of relation mark, annotation type of research objects and semantic similarity between test and train corpus. Then output the two relation mark triple list.

3.3 Test and evaluation

After about one year's work, we implemented the extraction system. We also carried out some experiments in which we successfully extracted thousand of research objects and their relations from scientific news. Figure 4 shows the result of a piece of scientific news' research object and relation extraction using the system.

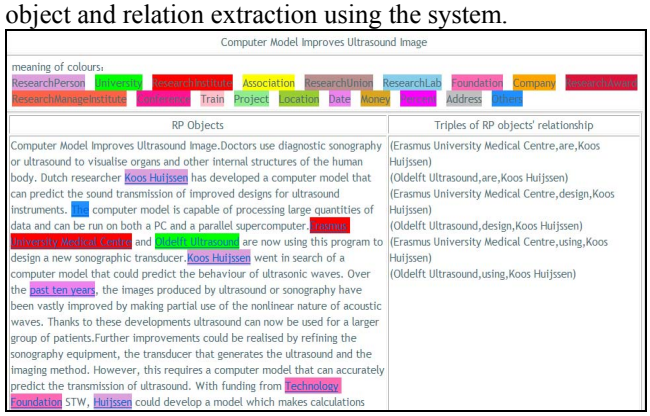

Figure 4. Research objects and relation triples extracted

To evaluation the system, we use it to deal with thousands of scientific news harvested from web. In this experiment, it processed 3945 pieces of news and average time cost is 4.26 seconds. Then we chose 1000 pieces randomly and divided them into 10 groups. Compared the extraction results by system and by hand, we compute the recall and precision of object and relation separately.

Table 1.Rresearch object and relation extraction evaluation (unit: %, R: Recall, P: Precisions)

group		1	2	3	4	5	6	7	8	9	10	AVG
object	R	79	85	81	77	79	82	78	80	81	84	80.6
	P	75	81	78	73	75	79	74	76	77	80	76.8
relation	R	35	34	39	34	35	38	41	32	37	34	35.9
	P	25	23	30	24	26	29	31	22	27	24	26.1

4 Methodology of research term extraction

Currently there are some term extraction tools available on Internet. After investigation and comparison, we design an improved term extraction method based on KEA according to web resources features.

4.1 Method of term extraction

KEA is an algorithm for extracting key phrases from text documents which based on following 4 feature values: *TF*IDF feature*, *First occurrence feature*, *Length feature* and *Node degree feature*[11]. After computing previous 4 features and probabilities, KEA gets a compound feature score for each candidate term and then output several terms present the main idea of document best based on the score.

However, according to experiment, KEA has good extraction show on well format periodical papers, but it is poor in web resources which have complex format and content component due to the following 4 aspects.

(1) The uneven content length of web resources causes low extraction quality. It's allowed to set the number of output terms (*e.g. each document output 20 terms with highest score*) in KEA. This method is effective on journal papers because of their long content length. But for those documents from web, it is hard to extract so many high qualified terms because they always have uncertain length.

(2) KEA adopts n-gram method to acquire candidate terms. So any phrase with high enough occurrence frequency may become candidate terms, including some phrases which have no any nominal component.

(3) There is no filter function module in KEA to filter those input corpus according to users' requirement. However, the aim of research term extraction is to support RP in certain domain. Therefore we need some filter module to filter domain web resources.

(4) Some of those terms outputting from KEA are common phrases without special domain features. These terms will affect the accuracy of following analyses tasks.

In view of web resources features, we made some improvements on KEA in following 4 aspects.

(1) Providing mechanism to acquire output terms based on probability score threshold. Content length of web documents always changes. Some documents only have several hundreds words while others' length is equal to common journal papers. If we adopt KEA directly and output same number of terms for each document, those short ones will output many low quality terms which will badly affect RP. So, we rebuild the output of KEA and users can set the max output number of terms and characteristic value threshold of output terms. After setting an appropriate characteristic value, those short documents will output less terms to keep term quality at a high level.

(2) Correcting the problem that KEA output terms without any noun word. We adopt Stanford Parser to parse sentences when original KEA process use n-gram method to acquire candidate terms. By parsing sentences, we get noun segment and then using them to filter candidate terms. The filter rule is whether candidate term has any noun segment. If has, we will keep it down to next step; if not, it means that the candidate term is not a real term and will be discarded (e.g. "*artificial"* will be discarded, and "*artificial intelligence"* will be kept down.)

(3) Using domain word list to filter input documents. Those unrelated domain web pages will affect the accuracy of PR tasks. Therefore, we filter the input documents before extracting. We get many key words from domain related journal papers published in recent years, evaluate the frequency of each one and remain the most frequent ones to form a domain core word list. During extraction, core words present frequency is evaluated, and the document is scored based on the frequency. Those documents with high score are considered that they have strong relation with target domain which need extracting while those with low score are considered that need not dealing with.

(4) Using stop words list to filter extract terms. After a series of filter and calculation, many terms that can express certain domain concepts are extracted by improved KEA method. But there are still some terms that can not express certain domain concepts because of their common nature or even no meaning at all. While continually extracting terms from domain related documents, we evaluate extracted terms and add those meaningless terms into stop words list. In the next term extraction round, the stop words list will be used for further improving the quality of extracted terms. (E.g. Terms *attribute, language* have no domain meaning and should be added into stop words list.)

The improved term extraction method we brought forward shows as follows.

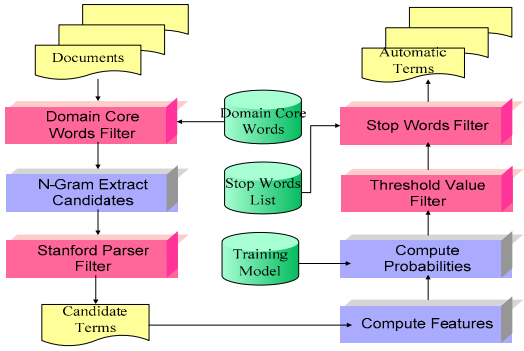

Figure 5. Improved term extraction method

（Note: Blue modules are original KEA components while red ones are new components designed ourselves.）

In detail, term extraction flow is composed of 6 steps.

(1) *Domain Core Words Filter*. The filter using domain core terms to evaluate documents content, and discard unrelated documents.

(2) *N-Gram Extract Candidates*. Using N-Gram method to segment sentences and generate candidate terms.

(3) *Stanford Parser Filter*. Stanford Parser Filter analyze each candidate term to judge if it has noun segment, then discard those terms without noun segment.

(4) *Compute Features*. Compute Features module compute TF*IDF feature, First occurrence feature and Length feature for each candidate term.

(5) *Compute Probabilities*. Compute Probabilities module uses Naïve Bayes algorithm to combine all features and generate a integrate score for each candidate term.

(6) *Threshold Value Filter*. Based on experience, we set a domain term score threshold. Those have scored higher than threshold is kept down.

(7)*Stop Words Filter*. Stop Words Filter module filter the candidate terms and those valid terms are output as final term extraction result.

4.2 Test and evaluation

For evaluation the improved KEA term extraction method, we select 10 de-noising web documents randomly, use KEA and improved one to extract terms separately. Table 2 shows the result of accuracy of each method.

Table 2. Evaluation result of term extraction (Unit:%)

Id / ACC	1	2	3	4	5	6	7	8	9	10	AVG
KEA	80	60	90	70	40	70	80	60	70	60	68
Improved KEA	100	78	90	75	80	90	90	80	100	78	86.1

From table 2 we can see that improved KEA method can adapt to the web resources features and always keep the higher extraction accuracy.

Take No.10 document for example: the original document content and extraction result using KEA and improved one separately are showed in Figure 6.

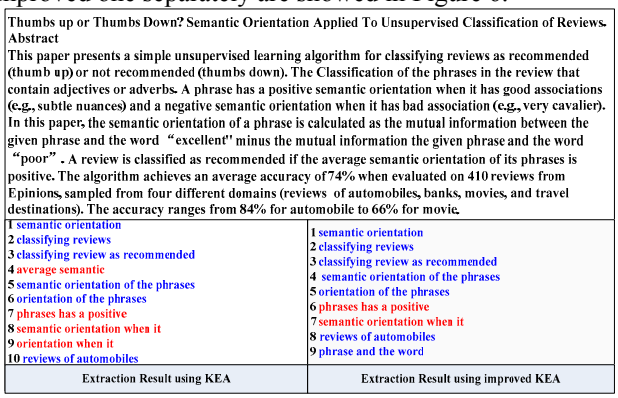

Figure 6. Original content of no.10 document

(Note: Terms marked with blue are correct according to artificial index terms while red ones mean false.)

As the figure show, the improved term extraction method can effectively reduce the extraction errors and output more useful terms.

5. Conclusions

This paper shows the method of extracting knowledge objects in scientific web resource for RP. Although we carried out some successful experiments and received good evaluation results, there is still a lot of work to do for its more efficient use. According to extraction results, we will adjust the existing rule set and algorithm; integrate more efficient approaches in knowledge object extraction. Besides, we will integrate the extraction system into our whole development trends of important scientific research institutions monitor and evaluation system.

Acknowledgements

This paper is supported by a project of National Key Technology R&D Program in the 11th Five year Plan of China named Science Monitoring and Evaluation based on Scientific Web Resources (2006BAH03B05)

References

[1] Porter, A. L., A. Kongthon, et al. "Research profiling: Improving the literature review". Scientometrics, 2002, 53(3): 351-370.

[2] Bragge, J., Sami R., et al. "Enriching Literature Reviews with Computer-Assisted Research Extraction". Case: Profiling Group Support Systems Research. In: Proceedings of the 40th Annual Hawaii International Conference on System Sciences, 2007, 243a.

[3] Bragge, J., Jan S. "Profiling Academic Research on Digital Games Using Text Extraction Tools". In: Proceedings of DiGRA 2007 Conference, 714-729.

[4] Hicks, D., G. Atlanta. "Global Research Competition Affects US Output". School of Public Policy, Georgia Institute of Technology, Atlanta, 2004, November 1.

[5] Nenadic, G., Irena S., et al. "Automatic discovery of term similarities using pattern mining" [EB/OL]. [2009-01-08]. http://acl.ldc.upenn.edu/coling2002/workshops/data/w05/w05-08.pdf.

[6] Zhao Jun. "Summary of Information Extraction Techniques Research" [EB/OL].[2008-8-30] http://159.226.21.7/file/ICCC2007.pdf.

[7] GATE Home [EB/OL].[2008-8-30]. http://gate.ac.uk/

[8] The Stanford Parser: A statistical parser [EB/OL].[2008-05-30]. http://nlp.stanford.edu/software/lex-parser.shtml.

[9] Athanasios T., Vangelis K., et al. "Learning of Semantic Relations between Ontology Concepts using Statistical Techniques" [EB/OL]. [2008-10-10]. http://www-ai.cs.tu-dortmund.de/HLIE08/slides/03-tegos-HLIE_08_Tegos.pdf.

[10] Marti A. H. "Automatic acquisition of hyponyms from large text corpora"[C]. In: Proceedings of the 14th International Conference on Computational Linguistics, 1992, 539-545.

[11] KEA Description [EB/OL]. [2009-3-1]. http://www.nzdl.org/Kea/description.html.