Glossary Construction of Scientific and Technical Term by using Wikipedia

Yong-Ki Kim and Jaesoo Kim*

NTIS Center, Korea Institute of Science and Technology Information, Daejeon - 305-806, Republic of Korea;
ykkim@kisti.re.kr, jaesoo@kisti.re.kr

Abstract
Among numerous encyclopedias, Wikipedia operated by the Wikimedia Foundation is the biggest as a single project among dictionary projects, and employs the concept of user participation to create contents. In this study, we propose an algorithm for building a glossary by using the Wikipedia API. The glossary is built by reflecting the latest trend to extract science and technology terms from the National R&D information, and using collocations of terms to create collocations of terms in order to use them as tags of terms. The glossary of terminology created as described above is used to know the latest trend, and used to create thesauruses of science and technology terms, for example, collocations of keywords in question.

Keywords: Glossaries, Scientific and Technical Information, Wikipedia

1. Introduction
The OECD Statistics Book published biannually for time series information since 1981 shows that investment in research and development by entire OECD member countries and non-member countries increased by 7% on the average between 2007 and 2011. In addition, the NTIS was built to enhance investment efficiency in national research and development and outcome productivity. The NTIS provides information about national R&D projects, for example, management of projects carried out with the national R&D budgets, R&D facility equipment and human resources. In particular, it supports service users with information through direct opinion exchange about R&D related contents including inquiries related to R&D and provided services. The open dictionary allowing users to share essential information and professional knowledge assists users to get information they want easily and more conveniently.

Meanwhile, Wikipedia operated by the Wikimedia Foundation, which is one of numerous dictionaries in Korea and other countries, is the biggest in size as a single project among dictionary projects, employs the concept of user participation and is evaluated as the most successful project. In particular, Wikipedia in Korean has grown to receive the contents of about 100,000 global encyclopedias from Daum and have the most documents in Korea. However, its reputation is not so good because of mistranslation of many documents. A limitation of Wikipedia is that it is a project built on line and made by everyone to freely use it. Apart from this, various institutions in Korea have built their own glossary of terminology. For example, the Defense Agency for Technology and Quality made the glossary of science and technology terminology for national defense, and the Research Information Center appointed by the National Research Foundation of Korea built glossaries of terminology for each field. It is necessary to use various types of glossaries of terminology related to science and technology to analyze, collocate and use the contents and structure thereof. A service is also required to examine, gather and collocate currently available resources related to the glossaries scattered across Korea and other countries.

This paper suggests a keyword extraction algorithm for national R&D information reflecting the latest trend to build a glossary of terminology. In this case, the
suggested algorithm measures associated similarity to build collocations in order to use them as a tag. The contents of the glossary of terminology are created by using the Wikipedia API.

Contributions of this study are described below.

- Researchers in universities and enterprise accumulate quality information by using the keywords for outcomes (projects, papers, patents) created through national R&D projects, and can understand the latest trend for planning and performing R&D.
- By building terminology by means of Wikipedia, NTIS users can understand the latest science and technology terminology, and terminology related to performing R&D.
- The open dictionary of Wikipedia allowing essential knowledge of a dictionary and professional knowledge to be shared supports users for the glossary of terminology to reflect their opinion.

The configuration of this study is as follows. Chapter 2 describes studies related to glossaries of terminology. Chapter 3 describes the background for building a glossary of terminology. Chapter 4 describes the algorithm for building the suggested glossary of terminology. Chapter 5 describes conclusion and future studies.

### 2. Related Work

#### 2.1 Encyclopedias in Other Countries

First, Jimmy Wales established a plan to make a new online encyclopedia with the wisdom of experts, semi-experts and ordinary people. He aimed to make the biggest knowledge storehouse by allowing all people to participate in editing the encyclopedia, not just by experts' participation. With this as a starting point, tens of thousands of people are currently participating in editing Wikipedia, which provides about 350 categories. While an encyclopedia edited mainly by an editor team is not ideal in the modern society in which just few clicks are required at home to get latest information, Wikipedia rapidly reflects latest information. Although Wikipedia has an issue involved in reliability of sources, conventional printed encyclopedia also involve low reliability issues in terms of sources. Nevertheless, Wikipedia approach is studied by many researchers.

Second, the Encyclopaedia Britannica which is the oldest among modern encyclopedias currently available have gone through overall revision at least 15 times so far since the first edition was published in Edinburgh, Scotland in 1768. The Encyclopaedia Britannica is globally based on viewpoint neutral, and the principle of organizing knowledge and information. It has continuously evolved in the best direction according to the request at the time concerned while keeping reliability of the value of knowledge. It has a dual structure of Macropaedia and Micropaedia to respond to the purpose of 'education' and 'reference', enables structural information search with cross reference, Propaedia and Index, and has complemented timeliness. At present, the Encyclopaedia Britannica is searching for a new scheme to cope with the digital age.

#### 2.2 Korean Glossary of Terminology

Encyclopedias providing glossaries of terminology of various fields are described below. The first is the Naver Encyclopedia with the most contents compared with other Korean encyclopedias. The Naver Encyclopedia built and operated is classified into the encyclopedia in which users can participate to create contents, and other various encyclopedias, for example, for animals, plants, children, and arts. The Naver Encyclopedia has at least 45 dictionaries with at least one million headwords in cooperation with 1000 dictionaries. The second is the NATE Encyclopedia based on the Encyclopaedia Britannica, which has built and operated additional contents, for example, the Academy of Korean Studies, and the Cultural Heritage Administration.

Second, glossaries of terminology built based on terms in various fields are described below. The first is the Naver Encyclopedia with the most contents compared with other Korean encyclopedias. The Naver Encyclopedia built and operated is classified into the encyclopedia in which users can participate to create contents, and other various encyclopedias, for example, for animals, plants, children, and arts. The Naver Encyclopedia has at least 45 dictionaries with at least one million headwords in cooperation with 1000 dictionaries. The second is the NATE Encyclopedia based on the Encyclopaedia Britannica, which has built and operated additional contents, for example, the Academy of Korean Studies, and the Cultural Heritage Administration.

Second, glossaries of terminology built based on terms in various fields are described below. First, the NATE Glossary of Terminology and the NAVER Glossary of Terminology handle most glossaries of terminology operated in portal sites. Second, an exemplary glossary of terminology built based on terms used in specific field is the IT Glossary of Terminology by the Telecommunications Technology Association, which is an online glossary of terminology supporting detailed search, for example, Exact, Like, Start, Intermediate and End. An exemplary glossary of computer terminology is the Glossary of Terminology that includes terminology of all CS fields. Each divided and specialized field builds and provides its own glossary of terminology.
3. National Science and Technology Information Service

3.1 Overview of National Science and Technology Information Service

The National Science and Technology Information Service (NTIS) gathers and manages national R&D project information carried out with national R&D budgets through government ministries of Korea and outcomes, for example, papers or patents obtained from the projects. The R&D information including projects, outcomes and research facility equipment are defined as 389 standard entry items, gathered and provided in real time. This process is performed once every year by investigating, analyzing and providing the project information carried out with national R&D budgets. The NTIS used information obtained through national R&D to build and provide various services. The services provided contribute to having the following effects.

- Improve national R&D investment efficiency by supporting the R&D management system of virtuous cycle.
- Efficiently manage national R&D projects through national R&D.
- Address current national questions and social issues by analyzing national R&D information from diverse viewpoint.
- Improve the efficiency of investment in research equipment by supporting the entire process of examining research equipment budgets.
- Establish a ministry-wide science and technology data governance system to maximize science and technology data management efficiency and availability.

3.2 Main Services of NTIS

The Project Management Service gathers all program and project information carried out with national R&D budgets to provide various services, for example, the service for searching for similar projects through project similarity, and the service for providing information about punishment taken while performing projects. The National R&D Participant Information Service is for the information of researchers who participate in national R&D programs, and the information is gathered by representative institutions of each government ministry. Therefore, if a researcher registers his/her own history information including paper outcomes just once to participate in R&D projects, it is not necessary to repeat the registration process. For the evaluator candidate recommendation, professional knowledge and technology of gathered research evaluators is checked to know whether they match the participant information of national R&D projects to exclude the involved evaluators. This can enhance evaluation objectivity and fairness. Researchers who participate in national R&D projects are given a researcher registration number for science and technology instead of their resident registration number for identifying individuals to protect personal information. The National R&D Outcome Information Service gathers outcomes including papers and patents obtained from R&D projects to comprehensively analyze and manage them. Outcomes are gathered by nine representative institutions for each type to gather and manage them in cooperation with the NTIS. The Research Facility Equipment Service provides research facility equipment information provided for national R&D projects, to allow researchers to get information about research facility equipment owned by each university and research institution. The Science and Technology Statistics Service provides statistics about science and technology including statistics about investment in science and technology, studies, outcomes and OECD statistics.

4. Algorithm for Building Science and Technology Terminology

It is necessary to submit a written project plan and results including papers, patents and research reports as outcomes after completing the project to be evaluated.
about the project plan in order to perform a project with national R&D budgets. In this case, it is necessary to enter keywords representing the concerned project or results. Because the keywords represent the details of a project performed, they allow diverse analysis of research trend, outcomes and research of patents. Therefore, a glossary of science and technology terminology is built by using keywords of projects, papers, patents and research reports created through R&D to extract terms related to science and technology that reflect the latest trend.

4.1 Construction Algorithm

Because the NTIS gathers and provides all science and technology information obtained from projects with the aforementioned national R&D budgets, the information is used to build an encyclopedia. The algorithm for extracting k terms in consideration of the latest trend is described below. The latest trend set (T) is composed of {keyword, number of appearances, collocated keyword set {collocated keyword, number of appearances}, existing set}. In this case, the number of appearances is the number of appearances within the period of latest search, the collocated keyword set is the collocate keywords and the number of appearances, and the existing collocation set is composed of flag TRUE or FALSE.

First, extract keywords from national R&D information including projects, papers and research reports. The extracted keywords are used to make a list of collocated keywords. Second, examine whether the keyword in question exists in the contents of an encyclopedia already built, or in the latest trend set. Third, if the contents of the encyclopedia already built include the keyword in question, but the latest trend keyword set does not have it, make a related tag with the collocated keyword set to include it in the latest trend set. Fourth, if the contents of the encyclopedia already built include the keyword in question, and the latest trend keyword set also has it, increment the keyword count in question by +1 and add the collocated keywords to the collocated keyword set. Fifth, if the contents of the encyclopedia already built do not include the keyword in question, but the latest trend keyword set has it, increment the keyword count in question by +1, and then add collocated keywords to the collocated keyword set. Sixth, if it is not included in the contents of the encyclopedia already built and the latest trend keyword set, include it in the latest trend keyword set and then initialize the keyword count in question to +1 and add the collocated keywords to the collocated keyword set. Seventh, calculate the latest trend value by using equation 1 when the latest trend set for the extracted keywords is completed. Lastly, return the k latest trend keywords.

Figure 2 shows the procedure of an algorithm for extracting science and technology terms. Figure 3 shows the relation between extracted contents and related keywords.

4.2 Algorithm for Building Terminology by using Wikipedia API

Wikipedia provides an API to allow users to use the contents of encyclopedias already built. The algorithm for building a glossary of terminology by using the Wikipedia API for extracted terms is described below.

First, call the contents of keywords to be built through the Wikipedia API. In this case, if the term is not in Wikipedia and there are substitute terms, bring the contents of substitute terms. Third, if the contents are brought from the substitute terms, register the extracted keyword as a collocation. Fourth, if Wikipedia does not include the keyword in question (or substitute term), recall the API instead of the keyword in question for the closest collocation. In this case, if Wikipedia does not include it, exclude the keyword in question from the term candidates for building a glossary. Equation 2 is about the term similarity by using the frequency of collocations. Lastly, save the keyword related to contents for the keyword called with the API as a tag.

Figure 2. Procedure of a Constructing Algorithm for Science and Technology Terms.
Figure 4 shows an interface for building a glossary of terminology by using the Wikipedia API. Figure 5 shows related contents and tag information linked with terms.

5. Conclusion

In this paper, algorithm for constructing glossary using Wikipedia API is proposed. Proposed algorithm reflects the latest trend from national R&D database, and measures the similarity of related keywords to use as tag. Content of keyword is collected from Wikipedia. This study suggests an algorithm for building a glossary of terminology by using the Wikipedia API. For building the glossary of terminology, the latest trend is reflected to extract terms related to science and technology from the national R&D information, and collocation similarity is used to build collocations used as a related tag of terms. This study contributes to checking the latest trend, and can be used for building thesauruses for science and technology terms including collocations of keywords in question. As a future work, it is required to study the efficiency of the algorithm for measuring and suggesting usability of tags made by the proposed algorithm.

6. Acknowledgment

This research was supported by Maximize the Value of National Science and Technology by Strengthen Sharing/Collaboration of National R&D Information funded by the Korea Institute of Science and Technology Information (KISTI).

7. References

1. Thomas JA. Managing the flow of technology: Technology transfer and the dissemination of technological information within the R&D organization. MIT Press Books. 1984; p. 1.
2. Coe DT, Helpman E. International R&D spillovers. European Economic Review. 1995; 39:859-87.
3. Available from: https://ko.wikipedia.org/
4. Remy M. Wikipedia: The Free Encyclopedia. Online Information Review. 2002; 26(6):434-5.
5. Leuf B, Ward C. The Wiki way: quick collaboration on the Web. Addison-Wesley; 2001.
6. Schaffert S. IkeWiki: A semantic wiki for collaborative knowledge management. Enabling Technologies: Infrastructure for Collaborative Enterprises. Manchester. 2006; p. 388-96.
7. Cole M. Using Wiki technology to support student engagement: Lessons from the trenches. Computers and Education. 2009; 52(1):141-6.
8. Wheeler S, Yeomans P, Wheeler D. The good, the bad and the wiki: Evaluating student-generated content for
collaborative learning. British journal of educational technology. 2008; 39(6):987-95.
9. Wang X. miRDB: a microRNA target prediction and functional annotation database with a wiki interface. A Publication of the RNA Society. 2008; 14(6):1012-7.
10. Wagner C. Wiki: A technology for conversational knowledge management and group collaboration. Communications of the Association for Information Systems. 2004; 13(1):265-89.
11. Engstrom ME. Jewett D. Collaborative learning the wiki way. TechTrends. 2005; 49(6):12-5.
12. Trentin G. Using a wiki to evaluate individual contribution to a collaborative learning project. Journal of Computer Assisted Learning. 2009; 25(1):43-55.
13. Notari M. How to use a Wiki in education: ‘Wiki based effective constructive learning’. Proceedings of the 2006 International Symposium on Wikis. 2006; p. 131-2.
14. Gabrilovich E. Markovitch S. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. IJCAI. 2007; 7:1606-1.
15. Bruns A. Blogs, Wikipedia, Second Life, and beyond: From production to prod usage. Peter Lang. 2008; p. 1-408.
16. Strube M, Ponzetto SP. WikiRelate! Computing semantic relatedness using Wikipedia. AAAI. 2006; 6:1419-24.
17. Cucerzan S. Large-Scale Named Entity Disambiguation Based on Wikipedia Data. EMNLP-CoNLL. 2007; 7:708-16.
18. Denning P, Horning J, Parnas D, Weinstein L. Wikipedia risks. Communications of the ACM. 2005; 48(12):152.
19. Volkel M, Krotzsch M, Vrandecic D, Haller H, Studer R. Semantic wikipedia. Proceedings of the 15th International Conference on World Wide Web. 2006; p. 585-94.
20. Suchanek FM, Kasneci G, Weikum G, Yago: A large ontology from wikipedia and wordnet. Web Semantics: Science, Services and Agents on the World Wide Web. 2008; 6(3):203-17.
21. Hoffart J, Suchanek FM, Berberich K, Weikum G. YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Proceedings of the Twenty-Third international joint conference on Artificial Intelligence. 2013; p. 3161-5.
22. Ye R. Encyclopaedic visions: Scientific dictionaries and enlightenment culture. Cambridge University Press; 2001.
23. Available from: http://www.ntis.go.kr/