Bioinformatics and Biosystems 2007, Vol. 2, No. 1, pp. 33-36

CiNet: GUI based Literature analysis tool using citation information

Sejun Lee, Kwang H. Lee

Dept. of BioSystems, KAIST

Abstract

Scientific literature is the most reliable and comprehensive source of knowledge for scientific and biomedical information. Citation information in the literature is also reliable source for linking between literatures. We proposed CiNet, a graphic user interface based tool that extracts the trend of the research using citation information. We can navigate related literatures and extract keywords from the linked literature using this tool. These extracted keywords will be helpful to researchers who want to survey the information.

Keywords: Citation, information extraction, Textmining

Introduction

The internet has basically changed the way we access the information. It is now a routine processto use search engines, such as Google, and to follow hyperlinks rather than reach for a reference book. [2] [4] Hyperlinking of related content in the internetmakes retrieval of information associative and extremely effective. However, these progressive changes in the way we approach information are not followed in the tools available for scientists. [5] Tracing hyperlink is also somewhat annoying task when we are searching cited paper. It needs more convenient way to access theliterature.

We need also whole picture of connected literature exploration in the internet. The World Wide Web, however, is so achieve an information resource, because it grows naturally in view of a subsequent retrieval of information. [1][3] Scientific publications contain also references to other publications; however, in the heterogeneous world of publishing houses and policies, it is difficult to make this reference network available for navigation; not all publications are electronically and freely available and a common standard for linking is not in sight.[5]We also have a problem to see whole picture of literature network.

To tackle theseproblems, we proposed CiNet which is literature navigator tool using citation information and text mining approach (Fig 1)

Corresponding Author:
Kwang H. Lee (E-mail: khlee@kaist.ac.kr)
This work was supported by the Korea Science and Engineering Foundation(KOSEF) through the National Research Lab. Program (No. 2005-01450), and the Korean Systems Biology Research Grant (2005-00343).

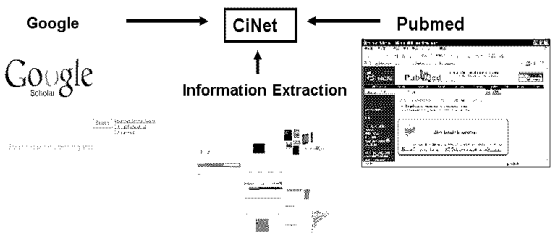

Figure 1. the System overview

Method

CiNet can be divided into two parts. The first one is GUI Interface part that can show the content to user. The second one is keywords extraction part. We implemented these two parts using Java programming

1. Construction CiNet with GUI Interface.

The system architecture of CiNet is shown figure 2. This system has two parts. The first part is indicated blue rectangle in fig 2, gathering information part. The other part is indicated red rectangle in fig 2, extracting information using textmining part.

1.1 Gathering information from Google

CiNet collected literature information which is citing inputted literature by entered literature name. Then, this program extracted literature title which is referred literature list from the Google Scholar. It retrieved this information in real-time by provided literature name.

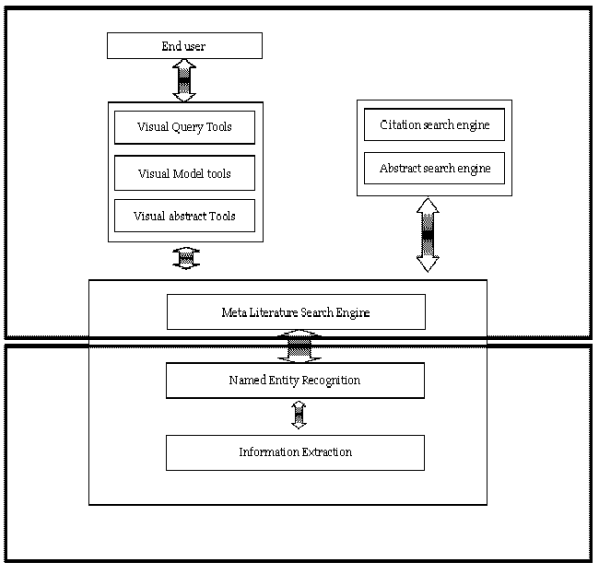

Figure 2.. the System architecture

1.2 Gathering abstracts from PubMed

We extracted abstract from PubMed using previously founding Google citation information.

2. Extracting keyword using text mining approach

All abstracts, which found previous step, were clustered the one integrated information. All abstracts in the cluster was then tokenized and filtered. The porter stemming algorithm is used here to normalize tokens (words). We produced a frequency list for cluster. Normally, this would be a word frequency list. Then we calculated scoring method.

2.1 Tokenizing the text

The step of tokenizing the text is the first step for analyzing the abstract. The purpose of tokenization is convertingthe text form to the word form. However, this tokenization result is dirty form such as including commas, periods, and so on. Furthermore, they can be represented various forms even they are same meaning.

2.2 Stop words elimination.

The second step for extracting keywords is stop words elimination. Stop words are those word which are so common that they are useless to index or search engines or other search indexes. In English, some obvious stop words would be "a", "of", "the", "I", "it", "you", and "and". These words occurred highly in the text. Therefore, we should deal with these words to find real keyword. We chose 391 stop words (Fig.3) which highly occurred but not meaningful in the text.

```
String stopList[] = {"a","about","above","across","afte
"back","be","became","because","become","becomes","beco
"due","during","each","eg","eight","either","eleven","e
"give","go","had","has","hasnt","have","he","hence","he
"made","many","may","me","meanwhile","might","mill","mi
"or","other","others","otherwise","our","ours","ourselv
"sometimes","somewhere","still","such","system","take",
"too","top","toward","towards","twelve","twenty","two",
"whom","whose","why","will","with","within","without","
```

Figure 3. Stop words

2.3 Porter stemming algorithm(M.F. Porter 1980)

We also used the porter stemming algorithm to deal with variation in the token. A stemming algorithm, or stemmer, is a computer program or algorithm for reducing inflected words to their stem, base or root form. We can deal with the words in fig. 4 as the same keyword using stemming algorithm.

Stem	CONNECT
variation	CONNECTED
variation	CONNECTING
variation	CONNECTION
variation	CONNECTIONS

Figure 4. Example for stemming algorithm

Frequently, the performance of an IR system will be improved if term groups such as this are conflated into a single term. This may be done by removal of the various suffixes -ED, -ING, -ION, IONS to leave the single term CONNECT (Fig.4). In addition, the suffix stripping process will reduce the total number of terms in the IR system, and hence reduce the size and complexity of the data in the system, which is always advantageous.

2.4 Searching keywords

We need to calculate the Expected values (E) according to the following formula:

$$E_i = \frac{N_i}{\sum_i N_i}$$

Figure 5. Significance calculation

In this formula, Niindicates number of occurrence word i in the cluster.

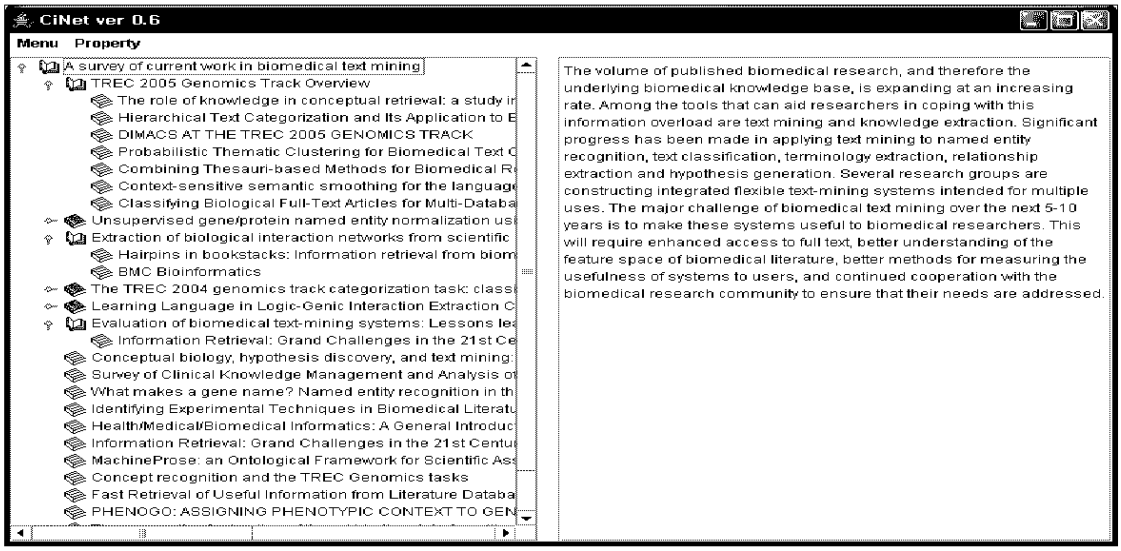

Figure 6. CiNet

Result

1. GUI Interface

Figure 6 shows CiNet interface. CiNet has two parts to show information for user. The left part represents literature name as tree structure. Linked literature means they are connected from citation. The right part is the abstract of selected literature. This directory structure is very familiar with user. Therefore, this program is also easy to deal with. Furthermore, representing abstracts with literature name give more convenience to user. This literature is also sorted from cited number, means how many articles cited in. Therefore, we can easily find that upper one is more important than below at the

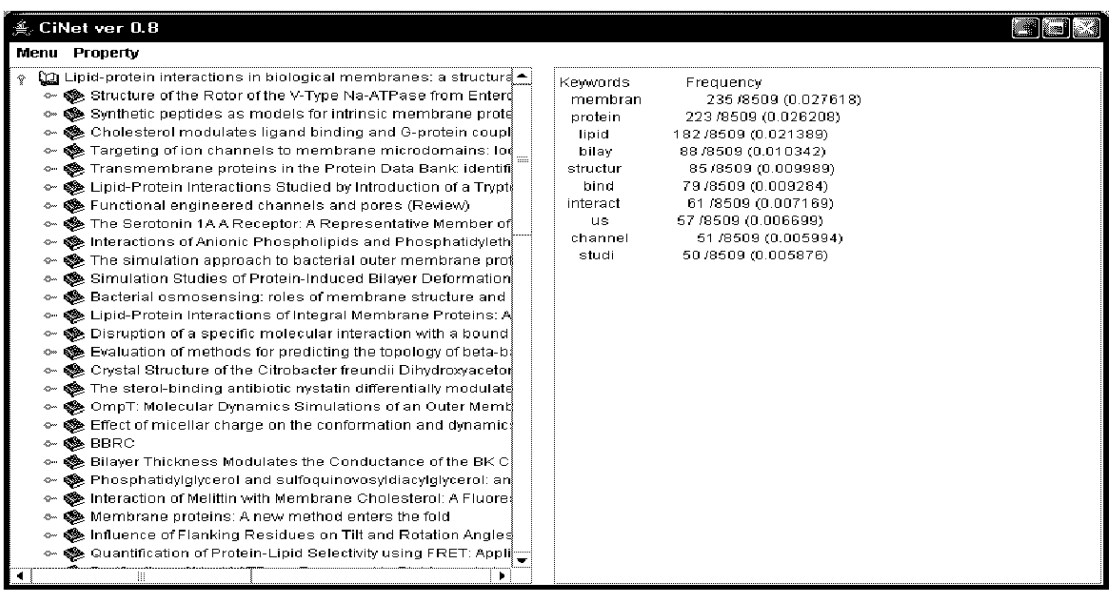

Figure 7. Keywords extraction for target literature(Lipid-protein interaction in biological membranes)

same level in the tree.

Keywords	Frequency
membrane	235 /8509 (0.027618)
protein	223 /8509 (0.026208)
lipid	182 /8509 (0.021389)
bilayer	88 /8509 (0.010342)
structure	85 /8509 (0.009989)
bind	79 /8509 (0.009284)
interact	61 /8509 (0.007169)
using	57 /8509 (0.006699)
channel	51 /8509 (0.005994)
studied	50 /8509 (0.005876)

Figure 8. Keywords from Figure 7

2. Keywords extraction

As you see the result Figure 8, top 10 ranked keywords are meaningful terms. Top 5 terms are related to the target literature. It other cases, this keywords extraction method is operated well.

Conclusion

In this paper we presented CiNet, a GUI tool that can be used to extract and visualize the information of literature from publications indexed by Google and PubMed. Distinguishing features of CiNet is that it includes its ability to generate tree structure and core abstract information using cited information based on a single query. Furthermore, keywords extraction function can provide kernel words toresearcher. It is useful for surveying. However, it does not provide enough information keyword itself. It will be better if we provide sentences, which contain keyword, retrieval tool.

There are also some limitations in this system. This system is not target-based system. Although we use the literature name, it is not specific target information. The keywords extraction is also just found the keyword from abstract cluster. However, researchers could be wanted to specific information. In this case, we should provide target-based system. This CiNet can provide general keywords, but it cannot deal with user's preference. However, the purpose of the CiNet is survey aided tool. In this point of view, CiNet is good enough to its objective.

References

[1] Barabasi, A.L. and Albert, R. (1999) Emergence of scaling in random networks. Science, 286, 509-512

[2] Henzinger, M. and Lawrence, S. (2004) Extracting knowledge from the World Wide Web. Proc Natl Acad Sci USA, 101, 5186-5191

[3] Huberman, B.A. and Adamic, L.A.(1999) Internet: Growth dynamics of the World Wide Web. Nature, 401,131

[4] Schatz, B.R. (1997) Information retrieval in digital libraries: bringing search to the net. Science, 275, 327-334

[5] Robert Hoffmann, and Alfonso Valencia (2005) Implementing the iHOP concept for navigation of biomedical literature. BIOINFORMATICS, vol.21 pages ii252-ii258

[6] M.F. Porter (1980) An algorithm forsuffix stripping. PROGRAM. 14. no.3 pp 130-137