Query Constraining Aspects of Knowledge
A Case Study

Ann-Marie Eklund
University of Gothenburg
Gothenburg, Sweden
ann-marie.eklund@svenska.gu.se

Abstract

In this paper we present a first analysis towards better understanding of the query constraining aspects of knowledge, as expressed in the most used public medical bibliographic database MEDLINE. Our results indicate that new terms occur, but also that traditional terms are replaced by more specific ones and decrease in use as major defining keywords, even though they are still used in abstracts. In other words, as knowledge, including terminology, evolve over time, queries and search methods will have to adapt to these changes to enable finding recent as well as older research papers in databases.

1 Introduction

When on-line databases are queried, answers are automatically derived from the contents of the database. For instance, MEDLINE (National Library of Medicine, 2010b) is a bibliographic database containing 18 million records from over 5000 biomedical journals, and for researchers in life science and medicine it is one of the most important on-line sources of new knowledge. Queries posted to the database are matched against bibliographic records of unstructured text (abstracts), titles and associated keywords. Each record aims at reflecting the knowledge of a given paper and its authors, and the best matching records are returned as answers to a query. In this case the answers are constrained by the bibliographic data. This data also has an impact on which terms would be the most efficient ones to use in a query. Since MEDLINE users, compared to users of general web search engines, are more persistent in their search for information and often reformulate their queries (Dogan et al., 2009), improved understanding of search behaviour may be of importance for automatic query optimisation. Another interesting question that rises in infodemiology, i.e. "the study of the determinants and distribution of health information" (Eysenbach, 2006), is how the data, hence the encapsulated knowledge, residing in a database, impacts the querying process.

In this paper we will present the first steps of an ongoing work towards better understanding of the constraining aspects of a database, by analysing a subset of MEDLINE records corresponding to the publications on the obesity-related protein adiponectin. This analysis indicates that new terms occur, but also that traditionally used terms are replaced by more specific ones and decrease in use as major defining keywords, even though they are still used in abstracts.

2 Materials and Methods

We used a corpus of 5851 MEDLINE (National Library of Medicine, 2010b) records (1993-2009) of bibliographic data, containing the term adiponectin[1] in title, abstract or keywords. We call this term an *anchor term* due to its role of defining the corpus. From each record we used title, abstract, year of publication and keywords. The keywords consist of Medical Subject Headings (MeSH) (National Library of Medicine, 2010a), which is NLM's controlled vocabulary thesaurus organised in a hierarchical structure.

The implementation[2] was done in Python using the Natural Language Toolkit (NLTK) (tokenization and lemmatization) and Biopython (data retrieval and management). The analysis of data was done using Microsoft Excel in combination with R

[1] We chose the term adiponectin because it is unambiguous and without synonyms, and due to its relatively new appearance in life science the corresponding corpus becomes rather easy to analyse.

[2] The program and result files can be obtained from the author on request.

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 279–282

(visualisation of data)[3]. Since this study is only an intitial analysis of the existing data, we have not utilised any other analysis methods than manual inspection of the data.

3 Results

In the adiponectin context, around 4500 different MeSH terms, or keywords, have been used since the first adiponectin paper in 1993. The abstracts contain around 20,000 different words, stopwords not included, and only a small part of the terms have been examined here. The emphasis in this section is on findings related to uses of the corpus anchor term (adiponectin), hyponyms of the words/terms used in the context of adiponectin, and the introduction of new terms over time.

3.1 Use of the anchor term

One interesting aspect of knowledge and its expression is if and when it becomes common, thereby more seldom explicitly stated in communication.

The first MEDLINE record containing the term *adiponectin* is from 1993, but before the year 2000 there are not many papers in MEDLINE mentioning *adiponectin* (figure 1, top). The number of papers containing *adiponectin* in title, abstract or keywords has increased every year since 1999, but more and more of the papers do not have *Adiponectin* as a keyword, (figure 1, bottom).

Hence, it seems like the use of the anchor term as a keyword has decreased over time.

3.2 Use of terms and their hyponyms

Since the MeSH keywords are hierarchically organised, it is possible to study if, and how, the use of more general (hypernym) and specific (hyponym) terms changes over time.

The percentage of papers having *Obesity* as a keyword decreased from around the year 2000. A corresponding decrease can be found in titles and abstracts, where we also have a percentage decrease in the use of the word *obese* (which is not by itself a MeSH term). The keyword *Obesity, Abdominal* is a hyponym of the term *Obesity* and can be found in the papers from the last two years. In the abstracts we see a frequent use of the word *abdominal* since 2003.

Other adiponectin related keywords, *Diabetes Mellitus* and its hyponym *Diabetes Mellitus, Type 2*, both first appeared in 2000 and both show a percentage decrease, although they increase in numbers. In titles and abstracts there is a percentage decrease for the term *diabetes*, but it is not as significant as for the keywords.

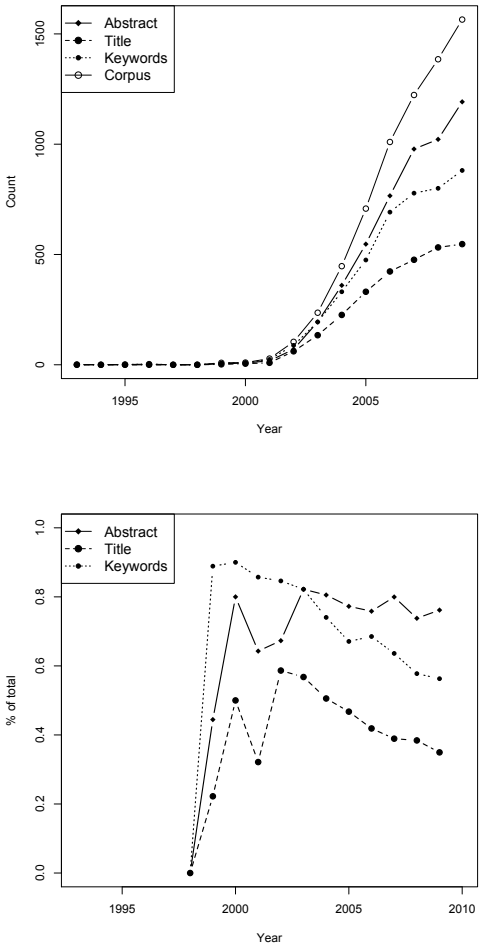

Figure 1: Number of MEDLINE records containing the term *adiponectin* in abstract, title or keywords (top), and the percentage of papers in the adiponectin corpus having the term in abstract, title and keywords respectively (bottom).

The keywords *Adipose Tissue* and *Adipocytes* were used in the first paper from 1993. They are both still in use as keywords, but there is a percentage decrease every year (figure 2). From 2007 *Adipocytes, White* and *Adipocytes, Brown* are being used as keywords. They are both hyponyms of the term *Adipocytes*. Similarly for *Adipose Tissue*, there are the hyponyms *Adipose Tissue, Brown*, first seen in 2002, and *Adipose Tissue,*

[3] nltk.org, biopython.org, r-project.org

White, which first occurred in 2006.

To conclude, by these examples we have seen indications of a shift over time in the use of traditionally used adiponectin related terms like *adipocytes*, *obesity* and *adipose tissue* towards the use of more specific terms (hyponyms).

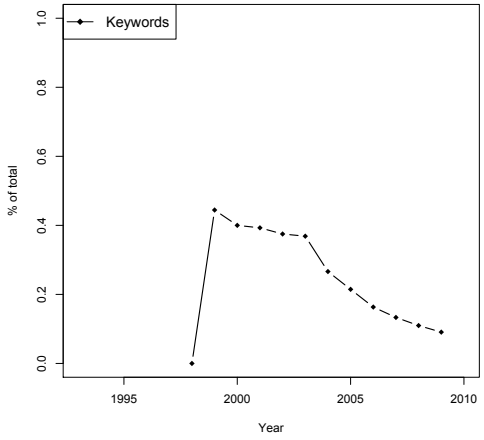

Figure 2: Percentage of papers indexed with keyword *Adipose Tissue*.

3.3 Use of new terms

If we assume that new knowledge and interests of a researcher are reflected in terms and keywords used in a paper, it is interesting to study if new words appear in our adiponectin corpus.

One interesting example of this is the increased use of words like *older*, *middle* and *aged* (figure 3) that we see in titles since their first occurrence in 2004. In abstracts *older* first appeared in 2003, and *middle* and *aged* in 2002. The keywords *Aged* and *Middle Aged* occurred for the first time in 1999, and since then both of them have been in frequent use. The keyword *Young Adult* is much used in 2009.

Another example is the plant related keywords. The keyword *Plant Extracts* has increased slightly since it was first used in 2005, and the keyword *Seeds* can also be found in a few papers every year since 2007 (*Seeds* is a descendant of *Plant Structures* or of *Food and Beverages* in the MeSH hierarchy). The last two years the keywords *Plant*, *Plant Stems*, *"Plants, Medicinal"* and *Plant Preparations* have appeared. In the last few years the words *plant* and *seed* have occurred mainly in abstracts, but also in a few titles.

Hence, by our analysis it is also possible to trace the occurrence of new terms, related to for instance age and plant concepts.

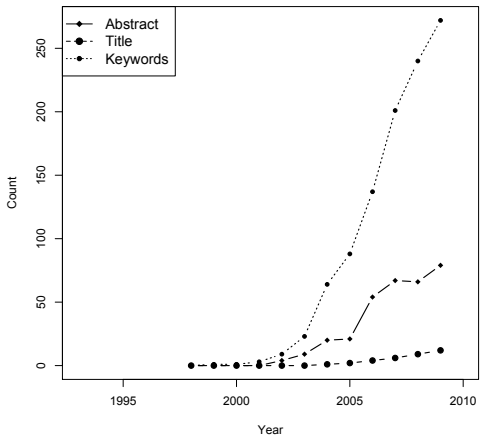

Figure 3: Number of occurrences of the term *aged*.

3.4 Other reflections

In addition to the topics above we have noted some variations of the term *adiponectin*.

Even in reviewed papers, misspellings of the anchor term adiponectin can be found in the abstracts, for example *adiponetin*, *adipnectin* and *adyponectin*. Another interesting aspect is that, based on the old term, new words have been created, e.g. *adiponectinaemia*, *adiponectinemic*, *hyperadiponectinemia*, which have all appeared after 2001.

4 Discussion

In this section we will discuss the uses of the corpus anchor term (*adiponectin*), hyponyms of terms and the introduction of new terms over time in the context of the results presented above. However, first we will elaborate on some aspects of the materials and methods used in this paper.

4.1 Materials and methods

Since MeSH is designed to reflect knowledge and use of terms in the field of biomedicine, a term may have been used in titles and abstracts for some time before it is introduced into the MeSH ontology and is available for use as an indexing term. Thereby, basing a trend analysis on only keywords may not reflect the actual use of terms, or expressed knowledge. We have not taken into ac-

count the year of introduction of a keyword into the MeSH ontology. When we find a keyword for the first time in the adiponectin context we do not know if this is the first time it is available as an indexing term or if it has been a MeSH term for some time. This may be slightly misleading when comparing the use of terms as keywords to their use in abstracts and titles.

In this study we only look at the term itself, and not at the term in combination with its hypernyms and hyponyms. If we took into account the terms above and below in the MeSH hierarchy, we might be able to see even more clear tendencies. For instance, by analysing the plant related terms and their hypo-/hypernyms together, we would add their different contributions, thereby making the new plant related aspect clearer.

In the light of the above limitations, our continued discussion will focus on the use of hyponyms and new terms.

4.2 Use of terms and hyponyms

In the examples in Results (section 3) we have seen indications of keywords becoming more specific, the annotations seem to have become more detailed, for example in the case of *Adipocytes* which decrease while its hyponyms *Adipocytes, White* and *Adipocytes, Brown* have started to be used as keywords. The use of more specific terms could indicate more detailed knowledge of a subject. To describe the new more detailed knowledge, new words may need to be used in the text, which may have led to the use of other more specific keywords to reflect that.

Another reason for the decrease in the use of for example terms like *Obesity* could be that obesity is already a given premise in this context and does not need to be stated explicitely anymore - terms become common knowledge, cf the above discussion on decreased use of the anchor term adiponectin.

4.3 Use of new terms

By studying the occurrence of new terms not used before in the adiponectin context, we find that terms related to completely new concepts appear. One example is the plant related terms, which correspond to an introduction of a new aspect into the research field. We can also see an increased age aspect, with terms like *Aged* and *Young Adult* being more and more common. New aspects like these often originate in the analysis of earlier study results, where new connections can be seen in the data and lead to new angles to study. When new terms appear, like the plant or age related terms in the adiponectin context, it could reflect new knowledge and new interests within the field. The increased use of plant related terms seen in the last few years could indicate an increasing interest in plants in medicine.

5 Conclusions

In the examples above, we have presented indications of a shift over time in the use of terms towards more specific terms (hyponyms), where the use of more specific terms could indicate a more detailed knowledge of a subject. There was also a decrease in the use of some keywords which are closely connected to the anchor term *adiponectin*. This decrease could indicate that the concepts described by these terms are already given in this context and that the concepts have become common knowledge. We have also seen examples of the appearance of new terms related to concepts not previously occurring in this context. This could be an indication of new knowledge being added to the existing one.

To summarise, in this paper we have tried to exemplify how the use of terms in bibliographic records changes over time, and how this may be related to the evolution of new knowledge. As a consequence, as knowledge evolve over time, queries and search methods will have to adapt to these changes, so that the search terms which are used reflect the actual contents of the papers in the database.

References

Rezarta Islamaj Dogan, G. Craig Murray, Aurelie Neveol, and Zhiyong Lu. 2009. Understanding PubMed user search behavior through log analysis. *Database (Oxford)*, 2009:bap018.

Gunther Eysenbach. 2006. Infodemiology: tracking flu-related searches on the web for syndromic surveillance. *AMIA Annu Symp Proc*, pages 244–248.

National Library of Medicine. 2010a. Fact sheet Medical Subject Headings (MeSH).

National Library of Medicine. 2010b. Fact sheet Medline.

ISSN 1736-6305 Vol. 11
http://hdl.handle.net/10062/16955