User Centred Networked Health Care
A. Moen et al. (Eds.)
IOS Press, 2011

doi:10.3233/978-1-60750-806-9-549

Medical Knowledge Evolution Query Constraining Aspects

Ann-Marie EKLUND[a,1]
[a]*Centre for Language Technology, Department of Swedish Language,*
University of Gothenburg, Sweden

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, possibly not surprising, that new terms occur, but also that traditional terms are replaced by more specific ones or even go out of use as they become common knowledge. Hence, as knowledge evolve over time, search methods may benefit from becoming more sensitive to knowledge expression, to enable finding new, as well as older, relevant database contents.

Keywords. Information Retrieval, NLP, Question Answering, Medical Informatics

1. Introduction

As presented by for instance Prier et al [1] and Bender et al [2], social media like Twitter and Facebook have changed the way people communicate health and disease related matters, by providing new ways of sharing experiences, and of seeking information and advice from others, both professionals and general public.

The appearance of social media has also brought renewed interest to questions regarding knowledge and language, for instance as expressed by Paul et al [3], "you are what you tweet". In other words, the way you express yourself reflects your knowledge and interests. Thereby, if your knowledge or interests change, your tweeting, or searching, changes too. One such change could be the appearance of terms expressing new or more specialised interests, also reflected in query and communication logs.

Studying interaction logs can provide increased understanding of people's knowledge and interests, but also of how these change over time in relation to, for instance, changes in society. One example is how internet query logs have been used for syndromic surveillance tracking flu related searches [4,5]. Hence, our knowledge will direct the way we search and share information. Other restricting aspects in the context of health and medical data are accessibility (patient record databases) and exponential data increase (medical bibliographic repositories), which will have an impact on querying behaviour.

To summarise, the way health related information is shared and retrieved is heavily influenced by aspects like knowledge of the relevant topics and how data is organised.

[1] Corresponding author: Ann-Marie Eklund, MA, University of Gothenburg, Box 200, 405 30 Gothenburg, Sweden; E-mail: Ann-Marie.Eklund@svenska.gu.se.

Thereby it is of importance in query optimisation and the development of future internet-based health support.

In this work we focus on one of these aspects, i.e. how knowledge is expressed in a medical bibliographic database (MEDLINE) and how it evolves over time. For instance, we show how more specific terms (hyponyms) are used over time at the same time as concepts become common knowledge and are not explicitly expressed. This complements studies of e.g. number of used terms, used terminology and search persistence [6,7,8,9].

2. Materials and Methods

We used a corpus of 5851 MEDLINE records (1993-2009), which in title, abstract or keywords contain the term adiponectin, herein called an *anchor term* due to its role of defining the corpus. 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 manageable for manual analysis. From each record we used title, abstract, year of publication and keywords. The keywords consist of Medical Subject Headings (MeSH)[2], which is NLM's controlled vocabulary thesaurus organised in a hierarchical structure.

The implementation[3] 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 (visualisation of data).[4] This study is an initial analysis of the data, performed by manual inspection of a few concepts known to be discussed in the context of adiponectin, but also ones that are new in this context. It focuses on when the terms first occurred and if their use increases or declines over time.

MeSH is designed to reflect knowledge and use of terms in the field of biomedicine, and a term may have been used in titles and abstracts for some time before it is available in the MeSH ontology for use as an indexing term. Thereby, a trend analysis based only on keywords may not reflect the actual use of terms, or expressed knowledge. We have not taken into account the year of introduction of a keyword into the MeSH ontology, which may be slightly misleading when comparing the use of terms as keywords to their use in abstracts and titles.

3. Results

In the adiponectin context, around 4500 different MeSH terms, or keywords, have been used since the first adiponectin paper in 1993, and 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, and the introduction of new terms over time.

[2] www.nlm.nih.gov/mesh

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

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

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 not many papers in MEDLINE mention *adiponectin* (Figure 1, left). 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, right).

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

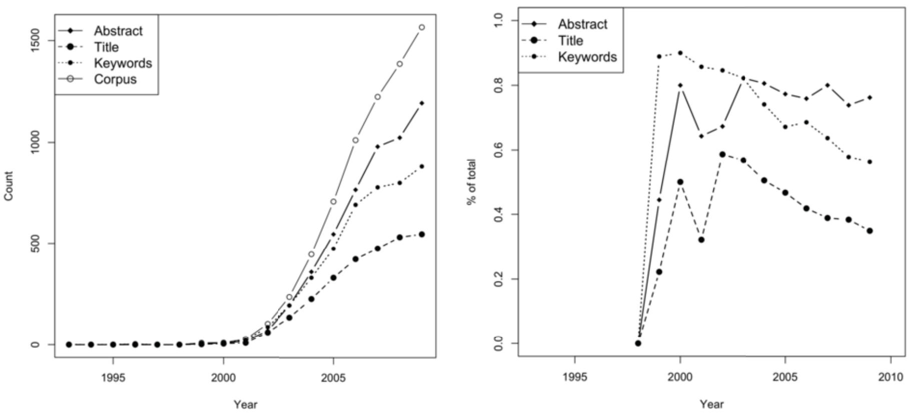

Figure 1. Number of MEDLINE records containing the term adiponectin in abstract, title or keywords (left), and the percentage of papers in the adiponectin corpus having the term in abstract, title and keywords respectively (right).

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 *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.

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. 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, White*, which first occurred in 2006.

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

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 the adiponectin context.

One example of this is the increased use of words like *older*, *middle* and *aged* 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*, *Plant Preparations* and *Plants, Medicinal* 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 more specialised study groups and alternative forms of treatment.

4. Discussion

4.1. Use of Terms and Hyponyms

In the examples in Results 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, described in the text by new terms not used before. This may have led to the use of 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 discussion in Results on the decreased use of the anchor term *adiponectin.*

4.2. 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 the results of earlier studies, 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 alternative treatments.

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), which 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.

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 may benefit from considering these changes, to make query terms match terms and keywords in the papers.

An approach for future investigation could be to make search algorithms "history aware", i.e. for a given search term, they could use its hypernyms to find older papers and its hyponyms to find more recent ones. This could be based on trend analysis of the occurrence of concepts/terms. A trend analysis could also note if given search terms decrease in use because they become common knowledge, and algorithms take this into account giving less weight to these terms when identifying relevant papers.

We have analysed only a limited bibliographic corpus and future work should address other corpora and if similar results can be found in other health domains [5].

References

[1] Prier K, Smith M, Giraud-Carrier C, Hanson C. Identifying Health-Related Topics on Twitter: An Exploration of Tobacco-Related Tweets as a Test Topic, *International Conference on Social Computing, Behavioral-Cultural Modeling, & Prediction (SBP 2011)*, 2011.

[2] Bender JL, Jimenez-Marroquin MC, Jaddad AR. Seeking support on Facebook: a content analysis of breast cancer groups, *J Med Internet Res* **13**(1) (2011).

[3] Paul MJ, Dredze M. You are what you tweet: Analyzing Twitter for public health, *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM 2011)*, 2011.

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

[5] Hulth A, Rydevik G, Linde A. Web queries as a source for syndromic surveillance, PLoS One **4**(2) (2009).

[6] Herskovic JR, Tanaka LY, Hersh W, Bernstam EV. A day in the life of PubMed: analysis of a typical day's query log, *J Am Med Inform Assoc* **14**(2) (2007), 212–220.

[7] Hoogendam A, Stalenhoef AFH, de Vries Robbé PF, Overbeke AJPM. Analysis of queries sent to PubMed at the point of care: observation of search behaviour in a medical teaching hospital, *BMC Med Inform Decis Mak* **8** (2008).

[8] Povnick RM, Zeng QT. Reformulation of consumer health queries with professional terminology: a pilot study, *J Med Internet Res* **6**(3) (2004).

[9] Dogan RI, Murray GC, Neveol A, Lu Z. Understanding PubMed user search behavior through log analysis, *Database* (2009).