A Graph-Based Analysis of Medical Queries of a Swedish Health Care Portal

Farnaz Moradi[1], **Ann-Marie Eklund**[2], **Dimitrios Kokkinakis**[2],
Tomas Olovsson[1], **Philippas Tsigas**[1]
[1]Computer Science and Engineering, Chalmers University of Technology, Sweden
[2]Språkbanken, Department of Swedish Language, University of Gothenburg, Sweden
{moradi,tomasol,tsigas}@chalmers.se[1]
{ann-marie.eklund,dimitrios.kokkinakis}@gu.se[2]

Abstract

Today web portals play an increasingly important role in health care allowing information seekers to learn about diseases and treatments, and to administrate their care. Therefore, it is important that the portals are able to support this process as well as possible. In this paper, we study the search logs of a public Swedish health portal to address the questions if health information seeking differs from other types of Internet search and if there is a potential for utilizing network analysis methods in combination with semantic annotation to gain insights into search behaviors. Using a semantic-based method and a graph-based analysis of word co-occurrences in queries, we show there is an overlap among the results indicating a potential role of these types of methods to gain insights and facilitate improved information search. In addition we show that samples, windows of a month, of search logs may be sufficient to obtain similar results as using larger windows. We also show that medical queries share the same structural properties found for other types of information searches, thereby indicating an ability to re-use existing analysis methods for this type of search data.

1 Introduction

Query logs which are obtained from search engines contain a wealth of information about the language used in the logs and the behavior of users. Searching for health and medical related information is quite common, and therefore analysis of query logs of medical websites can give us insight into the language being used and the information needs of the users in the medical domain.

In this study, we analyze 36 months of query logs from a Swedish health care portal, which provides health, disease, and medical information. On one hand, we perform a semantic enhancement on the queries to allow analysis of the language and the vocabulary which has been used in the queries. On the other hand, we perform a graph-based analysis of the queries, where a word co-occurrence graph is generated from the queries. In a word co-occurrence graph each node corresponds to a word and an edge exists between two words if they have co-occurred in the same query.

Our study reveals that a word co-occurrence graph generated from medical query logs has the same structural and temporal properties, i.e., small world properties and power law degree distribution, which has been observed for other types of networks generated from query logs and different types of real-world networks such as word association graphs. Therefore, the existing algorithms and data mining techniques can be applied directly for analysis of word co-occurrence graphs obtained from health search.

One of the widely studied structural properties of real-world networks is the communities in these networks. In this study, we apply a state-of-the-art local community detection algorithm on the word co-occurrence graph. A community detection algorithm can uncover a *graph community* which is a group of words that have co-occurred mostly with each other but not with the rest of the words in the network. The community detection algorithm used in this study is based on random walks on the graph and can find overlapping communities.

The communities of words, identified from the graph, are then compared with the communities of words obtained from a semantic analysis of the queries. In semantic enhancement, if a word or term in a query exists in medical oriented semantic resources, it is assigned a label. The words and terms which have co-occurred with these la-

Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi) @ EACL 2014, pages 2–10,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

bels are used to create a *semantic community*. We have compared the obtained semantic communities with the graph communities using a well-known similarity measure and observed that the communities identified from these two different approaches overlap. Moreover, we observed that the graph communities can cover the vast majority of the words in the queries while the semantic communities do not cover many words. Therefore, the graph-based analysis can be used to improve and complement the semantic analysis.

Furthermore, we study the effect of the time window lengths for analysis of log queries. Our goal is to investigate whether short snapshots of log queries also can be useful for this type of analysis, and how the increase in the size of the log files over time can affect the results.

The reminder of this paper is organized as follows. In Section 2 we review the related work. Section 3 presents the Swedish log corpus used for this study. Section 4 describes the semantic enhancement on the query logs. In Section 5 we describe the graph analysis methods. Section 6 summarizes our experimental results. Finally, Section 7 concludes our work.

2 Related Work

In this paper, we study the co-occurrence of words in medical queries and perform both a semantic and graph analysis to identify and compare the communities of related words. In this section, we briefly present a number of related works which deal with analysis of query logs.

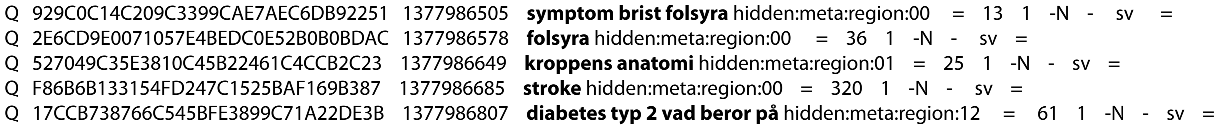
Query logs have been previously studied for identifying clusters of similar queries. In (Wen et al., 2001) a method was described for clustering similar queries using different notions of query distance, such as string matching of keywords. In (Baeza-Yates et al., 2004) clicked Web page information (terms in URLs) was used in order to create term-weight vector models for queries, and cosine similarity was used to calculate the similarity of two queries based on their vector representations.

Several previous works have also dealt with graph analysis of query logs. In (Baeza-Yates, 2007) several graph-based relations were described among queries based on different sources of information, such as words in the text of the query, clicked URL terms, clicks and session information. In (Herdagdelen et al., 2009) vector space models were compared, by embedding them in graphs, and graph random walk models in order to determine similarity between concepts, and showed that some random walk models can achieve results as good as or even better than the vector models. In (Gaillard and Gaume, 2011), it was shown that drawing clusters of synonyms in which pairs of nodes have a strong confluence is a strong indication of aiding two synonymy graphs accommodate each others' conflicting edges. Their work was a step for defining a similarity measure between graphs that is not based on edge-to-edge disagreement but rather on structural agreement.

3 Material - a Swedish Log Corpus

The Stockholm Health Care Guide, `http://www.vardguiden.se/`, is the official health information web site of the County of Stockholm, sponsored by the Stockholm County Council and used mostly by people living in the Stockholm aarea and provides information on diseases, health and health care. In January 2013 the Stockholm County Council reported that vardguiden.se had two million visitors per month. As of November 2013, vardguiden.se and another similar portal, 1177.se (which was a common web site for Swedish regions and counties, and the official national telephone number for health information and advice), are merged into one called 1177 Vårdguiden, sharing the same interface and search engine. The corpus data used in this study consists of the search queries for the period October 2010 to the end of September 2013. The data is provided by vardguiden.se, through an agreement with the company Euroling AB which provides indexing and searching functionality to vardguiden.se. We obtained 67 million queries in total, where 27 million are unique before any kind of normalization, and 2.2 million after case folding. Figure1 shows an example of a query log.

Information acquisition from query logs can be useful for several purposes and potential types of users, such as terminologists, infodemiologists, epidemiologists, medical data and web analysts, specialists in NLP technologies such as information retrieval and text mining, as well as, public officials in health and safety organizations. Analysis of web query logs can provide useful information regarding when and how users seek information for topics covered by the site (Bar-Ilan et



```
Q  929C0C14C209C3399CAE7AEC6DB92251  1377986505  symptom brist folsyra hidden:meta:region:00  =  13  1  -N  -  sv  =
Q  2E6CD9E0071057E4BEDC0E52B0B0BDAC  1377986578  folsyra hidden:meta:region:00  =  36  1  -N  -  sv  =
Q  527049C35E3810C45B22461C4CCB2C23  1377986649  kroppens anatomi hidden:meta:region:01  =  25  1  -N  -  sv  =
Q  F86B6B133154FD247C1525BAF169B387  1377986685  stroke hidden:meta:region:00  =  320  1  -N  -  sv  =
Q  17CCB738766C545BFE3899C71A22DE3B  1377986807  diabetes typ 2 vad beror på hidden:meta:region:12  =  61  1  -N  -  sv  =
```

Figure 1: Example queries. A query consist of (Q)uery, session ID, time stamp, search query, metadata, number of links returned, the batch ID of the visited link, (N)o spelling suggestions, Swedish search.

al., 2009). Such information can be used both for a general understanding of public health awareness and the information seeking patterns of users, and for optimizing search indexing, query completion and presentation of results for improved public health information. For an overview of some common applications and methods for log analysis see (Oliner et al., 2011).

Deeper mining into queries can reveal more important information about search engine users and their language use and also new information from the search requests; cf. (Medelyan, 2004). The basis for Search Analytics is made of different kinds of logs of search terms and presented and chosen results by web site users (Mat-Hassan and Levene, 2005). At a syntactic level queries may contain e.g., synonyms and hyponyms, and to be able to study patterns of search behavior at a more abstract level, we map the syntactic terms to semantic concepts. To our knowledge this is the first of its kind resource for Swedish and as such it can be used as a test bed for experimental work in understanding the breadth and depth of usage patterns, the properties of the resource and the challenges involved in working with such type of data. The only study we are aware of using Swedish log data, in the context of health-related information, is described by (Hulth et al., 2009). In their study, three million search logs from vardguiden.se (June 05 to June 07) were used for the purpose of influenza surveillance in Sweden, and seven symptoms, roughly corresponding to cough, sore throat, shortness of breath, coryza (head cold), fever, headache, myalgia (muscle pain) were studied.

4 Semantic Enhancement

Description of various corpus analytics that enables us to gain insights into the language used in the logs; e.g., terminology and general vocabulary provide, to a certain degree, an indication of the search strategies applied by the users of the web site service from where the logs are obtained. Findings can serve as background work that, e.g., can be incorporated in search engines or other web-based applications to personalize search results, provide specific site recommendations and suggest more precise search terms, e.g., by the automatic identification of laymen/novices or domain experts. The logs have been automatically annotated with two medically-oriented semantic resources (Kokkinakis, 2011) and a named entity recognizer (Kokkinakis, 2004). The semantic resources are the Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) and the National Repository for Medicinal Products (NPL, `http://www.lakemedelsverket.se/`)[1]. We perceive all these resources as highly complementary for our task since the Swedish SNOMED CT does not contain drug names and of course none of the two contain information about named entities.

4.1 SNOMED CT and NPL

SNOMED CT provides a common language that enables consistency in capturing, storing, retrieving, sharing and aggregating health data across specialties and sites of care. SNOMED CT provides codes and concept definitions for most clinical areas. SNOMED CT concepts are organized into 18 top-level hierarchies, such as Body Structure and Clinical Finding, each subdivided into several sub-hierarchies and contains around 280,000 terms. More detailed information about SNOMED CT can be found at the International Health Terminology Standards Development Organisation's web site, IHTSDO, at: `http://www.ihtsdo.org/snomed-ct/`.

The NPL is the official Swedish product registry for drugs and contains 11,250 entries. Every product in the registry contains metadata about

[1] Named entities have not been used for this study. However, we intend to use them in future studies. Nevertheless, the named entity annotation includes the ontological categories location, organization, person, time, and measure entities. Such entities can capture a wide range of entities searched by in such logs such as addresses to health care centers and various health care organizations.

its substance(s), names, dosages, producers and classifications, like prescription and Anatomical Therapeutic Chemical codes (ATC). For instance, for the question “missbruk st göranssjukhus” (“abuse st göran hospital”) from the query “Q \t C7ED234574EE24 \t 1326104437 \t missbruk st göranssjukhus meta:category:PageType;Article \t = \t 0 \t ...” (here “\t” signals a tab separation), we add three new tab-delimited columns (named entity label, SNOMED-CT, NPL or N/A if no match can be made) to each query. In this case, the three added columns for this particular query will get the labels “FUNCT-ENT”, “finding–32709003–missbruk” and “N/A” (no annotation), where the first stands for a FUNCTional-ENTity, the second for a finding category with concept-id “32709003” and “missbruk” as the recommended term.

4.2 Semantic Communities

We use the semantic labels obtained from the semantic enhancement to group words into communities. Communities can be used for getting insight into the language and the related words being used for medical search. The words which are matched with the same semantic label are clearly relevant to each other as they belong to the same semantic hierarchy. For each semantic label, we create a set of all the words in the queries which received this label. In other words, the words in queries that co-occurred with the same label are assumed to belong to the same community.

We have generated such communities only from SNOMED CT and NPL labels and refer to them as *semantic communities* in the rest of the paper. As an example, the community $\{$borrelia, serologiska, blodprover, test, serologisk, testning$\}$ was obtained from the queries which received the label “qualifier value–27377004–serologisk”.

5 Graph Analysis

Query log data can be modeled using different types of graphs (Baeza-Yates, 2007). In this study, we have generated a word co-occurrence graph, in which each node corresponds to a word and two nodes are connected with an edge if they have appeared in the same query. The generated graph is undirected and unweighted and has no multiedges. To generate the graph we have used the words as they appeared in the logs, i.e., we did not replace words with their synonyms, correct misspellings, or translate non-Swedish words to Swedish. For example, “eye”, “öga”, “ögat”, “ögon”, and “ögonen” appear as five different nodes in the graph but mean the same thing.

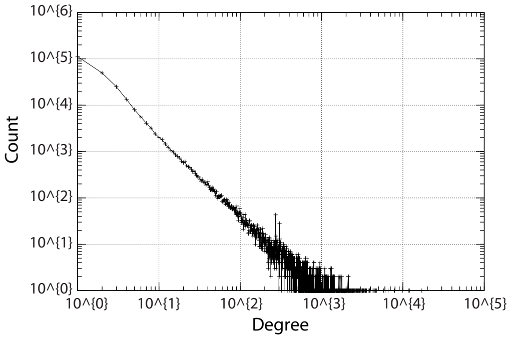


Figure 2: The degree distribution of the co-occurrence graph.

The graph $G(V, E)$ generated from the queries which contained two or more words has $|V| =$ 265,785 nodes and $|E| =$ 1,555,149 edges. The words in one-word queries which did not co-occur with any other words could not be considered for the graph analysis. The generated graph consists of 6,688 connected components. A connected component is a group of nodes where a path exists between any pair of them. The largest connected component of the graph, also known as giant connected component (GCC), contains around 95% of the nodes in the graph.

It was shown in (Ferrer i Cancho and Solé, 2001), that a graph generated from the co-occurrence of words in sentences in human languages, exhibit two structural properties that other types of complex networks have, i.e, the graph is a *small world* network and it has a *power-law degree distribution* (Barabási and Albert, 1999). Later studies on different types of word graphs have also been shown to follow the above properties. In this paper, we also show that a word co-occurrence graph generated from medical queries exhibits the same structural properties.

In small world networks, there is a short path connecting any pair of nodes in the GCC of the network. This property can be examined by calculating the *effective diameter* of the network (Leskovec et al., 2007). Small word networks also are highly clustered and therefore have a high *clustering coefficient* value. The effective diameter of our co-occurrence graph is 4.88, and it has an average clustering coefficient of 0.34. These values confirm that our word co-occurrence graph is a small world network.

Table 1: Structural properties of the word co-occurrence graph over time.

Time window	$\lvert V\rvert$	$\lvert E\rvert$	$\lvert V_{GCC}\rvert$	clustering coeff.	effective diameter
1 month	16,045	52,403	14,877	0.29	5.47
3 months	30,681	168,045	29,220	0.30	5.42
6 months	48,229	298,331	46,435	0.31	5.38
12 months	69,380	414,643	67,245	0.32	4.97
36 months	265,785	1,555,149	251,597	0.34	4.88

The degree distribution of the co-occurrence graph is shown in Figure 2. It can be seen that the degree distribution follows a power law distribution. This observation is similar to the observations presented by (Baeza-Yates and Tiberi, 2007) that almost all the measures of a graph generated from query log files follow power laws. Therefore, the user behavior in medical search does not seem different from general search behavior. In addition to networks of word relations, power law degree distributions have also been observed in social, information, and interaction networks where there are many nodes with low degrees and a few nodes with very high degrees (Clauset et al., 2009). The word with the highest degree in our graph is “barn” (child/children) which has 17,086 edges. Some other high-degree nodes are “sjukdom” (disease), “behandling” (treatment), “ont” (pain), “gravid” (pregnant), and “feber” (fever).

We have also looked into how the structural properties of the word co-occurrence graph change over time as the graph increases in size with an increasing number of queries. Table 1 summarizes the results. It can be seen that similar to many other networks, the diameter of the graph shrinks when more nodes become connected and its average clustering coefficient does not change much as the graph becomes larger.

Overall, the structural properties of the word co-occurrence graph are similar to many other real-world networks. Although it was shown in (Yang et al., 2011) that the queries and information needs of medical practitioners in accessing electronic health records are different from users of general search engines, our analysis reveals that there are similarities between information seeking of general users on health data and on general data. Therefore, the algorithms introduced for analysis of such networks can be directly deployed for analysis of word co-occurrence graphs.

5.1 Graph Community Detection

One of the widely studied structural properties of real-world networks is their community structure. A community, also known as a cluster, is defined as a group of nodes in a graph which have dense connections to each other, but have few connections to the rest of the nodes in the network. There have been numerous studies on the community structure of social and information networks and a variety of algorithms have been proposed for identifying the communities in these networks. A thorough overview of different types of community detection algorithms can be found in (Fortunato, 2010; Xie et al., 2013).

Community detection algorithms can be divided into global and local algorithms. The global algorithms require a global knowledge of the entire structure of the network to be able to find its communities. Therefore, these types of algorithms do not scale well for log analysis since query logs are usually very large and are continuously growing. The local algorithms, on the other hand, only require a partial knowledge of the network and therefore can identify network communities in parallel. However, the identified communities might not cover all the nodes in a network.

Moreover, community detection algorithms can be divided into overlapping and non-overlapping algorithms. Traditional partitioning and clustering algorithms typically divide the nodes in a network into disjoint communities. But in many real networks, a node can actually belong to more than one community. For example, in a social network, a user can belong to a community of family members, a community of friends, and a community of colleagues. In a co-occurrence graph, a symptom can co-occur with different types of diseases. Therefore, a community detection algorithm which can identify overlapping communities is more suitable for analysis of the graphs generated from search queries.

For the analysis of log queries, we have used a local overlapping community detection algorithm. This algorithm is a random walk-based algorithm which uses an approximation of a personalized PageRank (Andersen and Lang, 2006; Andersen et al., 2006) and is shown to perform well in detecting real communities in social and interaction networks (Yang and Leskovec, 2012). The algorithm starts from a seed node and expands the seed into a community until a scoring function is optimized. One of the widely used functions for community detection is *conductance*. The conductance of a community C in a graph $G(V, E)$ is defined as $\phi(C) = \frac{\overline{m}(C)}{\min(vol(C), vol(V \setminus C))}$, where $\overline{m}(C)$ is the number of inter-cluster edges and $vol(C) = \sum_{v \in C} deg(v)$ is the volume of a community and corresponds to the sum of the degree of all the nodes in the community. The lower the conductance of a community, the better quality the community has. The complexity of this algorithm is independent of the size of the network and only depends on the size of the target communities.

6 Experimental Results

In this section we present our experimental results and discuss the possible applications for graph-based analysis of medical data.

6.1 Semantic and Graph Analysis

From the semantic enhancement, we have generated 16,427 unique semantic communities which cover less than 11% of the nodes in the network. This means that, the majority of the queries in the network did not contain words that match the medical concepts provided by of SNOMED CT and NPL. This observation suggests that a semantic enhancement of queries on its own is not adequate for understanding the relations between all the words used in medical search.

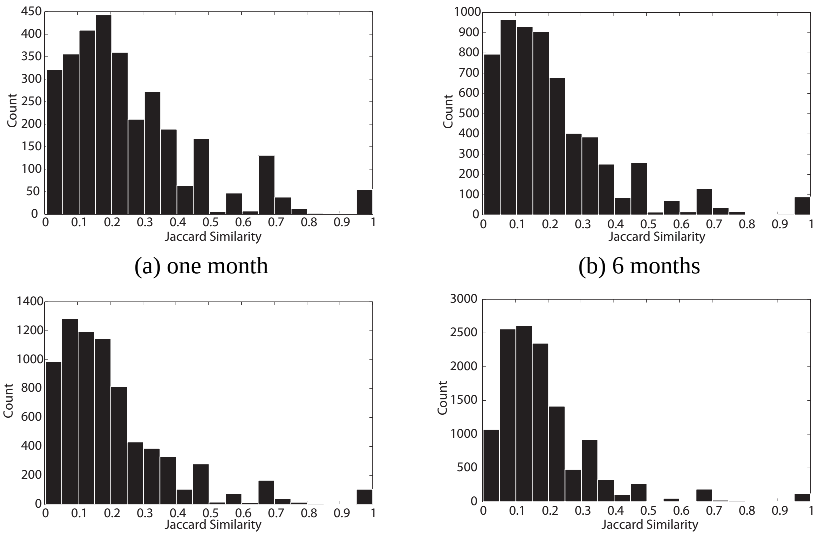
For the graph analysis, we have used the local overlapping community detection algorithm of (Yang and Leskovec, 2012) to identify the communities from the co-occurrence graph generated from the complete query logs. The algorithm identified 107,765 unique communities in the GCC of the graph with average conductance 0.74. This shows that the communities are not well separated from each other and that there are many edges between distinct communities. Moreover, the identified communities cover 93% of the nodes in the network which means that the graph analysis is more suitable for the study of the relations between the words than the semantic analysis.

The semantic communities and the graph communities are both dependent on the co-occurrence of words in queries, but identify communities differently. The semantic method places the nodes which belong to the same semantic hierarchy together with the words that co-occurred with them in the same community. However, the graph-based method places the words based on the structure of the generated network in the communities.

We have compared and calculated the similarity between the graph communities and the semantic communities using the *jaccard index* which is defined as $JI(C, S) = \frac{|C \cap S|}{|C \cup S|}$. The jaccard index shows the normalized size of the overlap between a graph community C and a semantic community S. Similarity functions, including Jaccard, have been used before for measuring the distance of two different queries. In this study we use similarity to assess the similarity of communities of words obtained from the two distinct methods.

We have compared each semantic community with all the graph communities and show the similarity distribution in Figure 3. It can be seen that the majority of the communities partially overlap. As an example, from the word “tandsjukdom” (dental disease) as the seed, we identified the graph community {tandsjukdom, licken, munhåleproblem, rubev, emalj, tändernaamelin, hypopla, permanentatänder, lixhen, hypoplazy, hipoplasy, hypoplazi, bortnött, hipoplazy}. From the semantic enhancement, “tandsjukdom” and “tandsjukdomar” both have received semantic label “disorder–234947003–tandsjukdom”. From the queries which received this label we have generated the semantic community {tandsjukdom, emalj, olika, vanligaste, tandsjukdomar, licken, plack, ovanliga}. The similarity of these communities is low, i.e., 0.16, however, they both contain the words which are clearly relevant to teeth and dental diseases.

As another example, “osteoklast” and “osteoklaster” both receive the semantic label “cell–27770000–osteoklast”. From the graph analysis, we have found {osteoklaster, osteoblster, osteocyter, osteoblaster} as a community with “osteoklaster” as the seed. We have also obtained the semantic community {osteoblaster, osteoklast, osteoporos, osteocyter, benskörhet, osteoklaster, osteoblster}. In this example, the graph community is a subset of the semantic community, and their similarity is 0.57. The above examples suggest that a graph-based analysis of medical queries can be used to complement the semantic analysis.



(a) one month

(b) 6 months

(c) 12 months

(d) 36 months

Figure 3: The distributions of jaccard similarity of semantic-based and graph-based communities.

6.2 Frequent Co-Occurrence Analysis

In the query logs, we observed that there are many misspellings, meaningless words, etc. In order to clear the dataset, it is common in different studies of log files, to filter out queries which appeared less frequently. By removing such queries, we can dramatically reduce the number of such words.

In this study, we have generated another graph from the words which co-occurred frequently in different queries. We have only considered words that co-occurred five times or more, and the graph contains 32,449 nodes and 217,320 edges, with average clustering coefficient of 0.29 and effective diameter of 5.66.

In the GCC of this graph we found 22,890 graph communities with average conductance of 0.65 and coverage of 95%. Moreover, we have also used the words which co-occurred at least five times to generate the semantic communities. The similarity of these communities with graph communities using jaccard similarity was 0.16 in average which is slightly lower than when no filtering was used. Overall, our observations suggest that filtering can be used to reduce the noise in the datasets and allow us to perform a faster analysis on a smaller graph.

6.3 Time Window Analysis

Another property which we have empirically studied in this paper is the effect of time window length during which the queries are analyzed. We have observed that, in average, more than 31% of the nodes and 12% of the edges have reappeared in each month compared to their previous month. This suggests that the search content changes over time perhaps depending on the changes in the monthly or seasonal information requirements of the users. It also means that over time the size of the word co-occurrence graph increases (see Table 1), and since in each month new co-occurrences shape, the graph becomes more and more connected. Therefore, when the time window is long, the analysis requires more time and the identified communities do not have good conductance. When the time window is short, the small size of the graph speeds up the analysis but might affect the analysis result. In this section we investigate the effect of time window length on our analysis.

We started by setting the time window length to one month. From the queries which were observed during each month, we generated a co-occurrence graph and identified the graph communities and the semantic communities. As presented in Section 5, the structural properties of a graph generated from one month are quite similar to that of the complete graph. We have also observed that the average conductance of the communities identified by the community detection algorithm is around 0.5 which is lower than when the complete graph was used. This means that the communities in the graphs generated from one month of queries have better quality since they have fewer connections to the rest of the graph.

We observed that the similarities between graph communities and semantic communities are higher when a one-month window is used (in average 0.26). By increasing the length of the time window from one to three, six, twelve, and thirty-six months, we observed a reduction in the similarities (in average 0.23, 0.22, 0.21, and 0.19, respectively). The similarity distributions are shown in Figure 3. It seems that with more queries over time, more words get connected and it becomes more difficult to identify good communities. Therefore, using short time windows can improve the quality of the analysis. Moreover, analysis of different time windows can also shed light on how the word relations and user requirements are affected by the months or seasons of the year.

6.4 Discussion

Our empirical analysis of a large-scale query log of medical related search presented in this paper can be used to improve our knowledge of the terminology and general vocabulary, as well as the search strategies of the users. In addition to providing a background for language analysis, a potential application for community detection could be to provide better spelling suggestions to users. We have observed that there are communities with very low conductance which contain a number of words which seem to correspond to guessing attempts to find a correct spelling, e.g., {shoulder, froozen, frosen, cholder, sholder, fingers, frozen, scholder, shulder, schoulder, shoulders}. The low conductance of the community means that the community is very isolated and has very few edges outside it and therefore it can easily be cut from the graph. Therefore, the community detection can be used for identifying such cases.

Another potential application of our graph analysis method is to provide recommendations and suggest more precise search terms based on the words that appear in the same community as the keywords entered by the users. For example, since the communities can overlap, each word can belong to more than one graph community or semantic community. We observed that in average, in the complete graph (generated from 36 months of logs), each word belongs to 3.8 unique graph communities and 3.6 semantic communities. It means that a word which can be related to multiple groups of words or have different meanings, can belong to several communities. This knowledge can potentially be used to provide suggestions to the users and help them to select the intended meaning and therefore reducing the ambiguity in the searched queries.

Overall, in this paper, we have presented a promising approach for analysis of medical queries using co-occurrence graphs. As a future work, the following improvements could be of interest for complementing our empirical study:

- Representing different variations of the words with only a single node in the graph, e.g., “öga” for “ögat”, and “ögon”.
- Filtering out the non-medical related words such as person and location entities from the queries based on the semantic enhancement with name entities from NER. Overall, more than 136,000 queries contained a person name entity, and around 127,000 contained a place entity.
- Filtering out high frequency words/terms which do not have medical significance, e.g., “olika” (different).

7 Conclusions

Our analysis of a large-scale medical query log corpus is the first step towards understanding the language and the word relations in health/medical related queries. We have performed a semantic enhancement of queries based on medically related semantic resources to find the communities of words which have co-occurred with a semantic label. We have also performed a graph-based analysis of the word co-occurrences and have shown that since a word co-occurrence graph has similar structural properties to many types of real-world networks, existing algorithms for network analysis can be deployed for our study. We then have used a random walk-based community detection algorithm in order to identify communities of words in our graph. Our empirical results show that the communities identified from the semantic analysis and the graph analysis overlap, however the graph-based analysis can identify many more communities and achieves much higher coverage of the words in the queries. Therefore, the graph-based analysis can be used in order to improve and complement the semantic analysis. Our experiments also show that short time window lengths for analysis of query logs, such as a month, would suffice for graph-based analysis of medical queries.

8 Acknowledgments

We are thankful to Adam Blomberg, CTO, Euroling AB for providing the log data. We are also thankful for the support by the Centre for Language Technology (`http://clt.gu.se`).

References

Reid Andersen and Kevin Lang. 2006. Communities from seed sets. In *Proceedings of the 15th international conference on World Wide Web - WWW '06*, page 223. ACM Press.

Reid Andersen, Fan Chung, and Kevin Lang. 2006. Local Graph Partitioning using PageRank Vectors. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*, pages 475–486. IEEE.

Ricardo Baeza-Yates and Alessandro Tiberi. 2007. Extracting semantic relations from query logs. *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '07*, page 76.

Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. 2004. Query Clustering for Boosting Web Page Ranking. In *Advances in Web Intelligence*, volume 3034, pages 164–175. Springer.

Ricardo Baeza-Yates. 2007. Graphs from Search Engine Queries. In *Theory and Practice of Computer Science*, volume 4362, pages 1–8. Springer.

Judit Bar-Ilan, Zheng Zhu, and Mark Levene. 2009. Topic-specific analysis of search queries. In *Proceedings of the 2009 workshop on Web Search Click Data - WSCD '09*, pages 35–42. ACM Press.

A.L. Barabási and R. Albert. 1999. Emergence of Scaling in Random Networks. *Science*, 286(5439):509.

Aaron Clauset, Cosma Rohilla Shalizi, and M. E. J. Newman. 2009. Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703, November.

R Ferrer i Cancho and R V Solé. 2001. The small world of human language. *Proceedings. Biological sciences / The Royal Society*, 268(1482):2261–5, November.

Santo Fortunato. 2010. Community detection in graphs. *Physics Reports*, 486(3-5):75–174, February.

Benoît Gaillard and Bruno Gaume. 2011. Invariants and Variability of Synonymy Networks : Self Mediated Agreement by Confluence. In *Proceedings of the TextGraphs-6 Workshop (Graph-based Algorithms for Natural Language Processing)*, pages 15–23.

Amaç Herdagdelen, Katrin Erk, and Marco Baroni. 2009. Measuring semantic relatedness with vector space models and random walks. In *In Proceedings of the TextGraphs-4 (Graph-based Methods for Natural Language Processing)*, pages 50–53.

Anette Hulth, Gustaf Rydevik, and Annika Linde. 2009. Web queries as a source for syndromic surveillance. *PloS one*, 4(2):e4378, January.

Dimitrios Kokkinakis. 2004. Reducing the Effect of Name Explosion. In *In Proceedings of the LREC Workshop: Beyond Named Entity Recognition, Semantic Labelling for NLP tasks. Fourth Language Resources and Evaluation Conference (LREC)*, pages 1–6.

Dimitrios Kokkinakis. 2011. What is the Coverage of SNOMED CTon Scientific Medical Corpora? *MIE: XXIII International Conference of the European Federation for Medical Informatics. Studies in Health Technology and Informatics*, 169:814 – 818.

Jure Leskovec, Jon Kleinberg, and Christos Faloutsos. 2007. Graph Evolution: Densification and Shrinking Diameters. *ACM Transactions on Knowledge Discovery from Data*, 1(1):2–es, March.

Mazlita Mat-Hassan and Mark Levene. 2005. Associating search and navigation behavior through log analysis. *Journal of the American Society for Information Science and Technology*, 56(9):913–934, July.

Olena Medelyan. 2004. Why Not Use Query Logs As Corpora? In *Proceedings of the Ninth ESSLLI Student Session*, pages 1–10.

Adam Oliner, U C Berkeley, and Archana Ganapathi. 2011. Advances and Challenges in Log Analysis Logs contain a wealth of information for help in managing systems . *Queue - Log Analysis*, pages 1–11.

Ji-rong Wen, Jian-yun Nie, and Hong-Jiang Zhang. 2001. Clustering user queries of a search engine. In *Proceedings of the tenth international conference on World Wide Web - WWW '01*, pages 162–168. ACM Press.

Jierui Xie, S Kelley, and BK Szymanski. 2013. Overlapping community detection in networks: the state of the art and comparative study. *ACM Computing Surveys*, 45(4).

Jaewon Yang and Jure Leskovec. 2012. Defining and evaluating network communities based on ground-truth. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pages 1–8.

Lei Yang, Qiaozhu Mei, Kai Zheng, and David a Hanauer. 2011. Query log analysis of an electronic health record search engine. *AMIA ... Annual Symposium proceedings / AMIA Symposium. AMIA Symposium*, 2011:915–24, January.