

# Automatic Identification of Key Concepts in Large PubMed Retrievals

Lana Yeganova, Vahan Grigoryan, Won Kim, W. John Wilbur

National Library of Medicine, National Institutes of Health

9000 Rockville Pike, Bethesda MD 20814

{yeganova, wonkim, wilbur}@mail.nih.gov, grigoryan\_vahan@bah.com

## Abstract

PubMed queries frequently retrieve thousands of documents making it very challenging for a user to identify information of interest. In this paper we propose a method for automatically identifying central concepts in large PubMed retrievals. The centrality of a concept is modeled using the hypergeometric distribution. Retrieved documents are grouped by concept, which can help users navigate the retrieval. We test our method on five datasets, each representing a medical condition.

## Introduction

MEDLINE® is a collection of approximately 21 million bibliographic records as of June 2012. It has grown exponentially and doubled in size within the last decade. With such abundance of information many queries retrieve thousands of documents and it is becoming increasingly difficult for users to browse the results which are presented in the reversed chronological order and find the information most relevant to their topic of interest.

The goal of this study is to consider a set of documents on a single topic and automatically identify key concepts in that set. Once the concepts are identified, documents can be grouped by concepts. By looking at these concepts users can get a perspective view of the topics discussed in a given set and navigate them faster and easier for obtaining relevant results.

In this paper we model the centrality of a concept in a set of documents  $S$  using the hypergeometric probability distribution. We compute the frequency  $f$  of a given concept in  $S$  and calculate the probability of observing that concept  $f$  or more times. When a concept is observed in a set more frequently than expected by chance, the resulting probability may be very low, indicating that the concept is pertinent to the set of documents  $S$ .

In the following section we describe how we process Medline to compile a comprehensive list of multiword phrases. We then consider a set of documents retrieved from PubMed in response to a query and process that set to

identify concepts that appear in these documents. We score the concepts and choose the highest scoring fifty. Then the optimal coverage algorithm is applied to choose the most diverse ten concepts that maximize the coverage of the set.

## Identifying Key Concepts

Biological concepts are frequently expressed in terms of phrases. Thus, to identify key concepts in a set of documents it is important to detect phrases, rather than single words. Here we describe how we process Medline to compile a comprehensive list of all phrases.

### Extracting Phrases from Medline

We read entire Medline and collect all multiword text strings bounded by punctuation or stop words. We then select text strings that appear at least twice on their own and at least five times as part of longer segments in Medline. This process results in 8.3 million unique text strings, which we will refer to as phrases. We further analyze these phrases to identify string variants. We consider two strings as variants if one string can be derived from another by word order permutation, stemming, or by consulting UMLS®. A detailed explanation of UMLS can be found at <http://www.nlm.nih.gov/research/umls/>. String variants are clustered into synonymy classes producing 1.3 million classes involving 3.2 million phrases of the original 8.3 million phrases. An example of a synonymy class consisting of 16 synonymous multiword strings is presented in Table 1.

We further consider synonymy classes and choose one of the phrases as a representative for the group. We use the method developed by (Kim et al. 2012) for detecting well-formed biomedical phrases which scores the phrases in each synonymy class. We select the highest scoring phrase as a representative for that class.

### Selecting Central Concepts

In this section we consider a set of documents  $S$  and identify concepts that appear to be central for the set. We

use phrases identified in the previous section and MeSH® terms as potential concepts. A detailed explanation of MeSH can be found at <http://www.nlm.nih.gov/mesh/>.

Let  $N_s$  be the size of the document set  $S$  and  $N$  be the size of Medline. For every phrase/MeSH term we compute two values:  $N_{st}$ , frequency of the phrase/MeSH term in  $S$ , and  $N_t$ , frequency of the phrase/MeSH term in all of Medline. A random variable  $Y$  representing the frequency of a phrase/MeSH term in the set  $S$  is a hypergeometric random variable with parameters  $N_s$ ,  $N_t$  and  $N$  (Larson 1982). The probability function of  $Y$  is:

$$P(y) = \binom{N_t}{y} \binom{N - N_t}{N_s - y} / \binom{N}{N_s}$$

We now compute the p-value, i.e. the probability of observed or more extreme frequency arising by chance as:

$$\text{p-value} = \sum_{y=N_{st}}^{\min(N_s, N_t)} P(y)$$

and use the negative  $\log$  of p-value as a measure of significance of the phrase/MeSH term relative to set  $S$ . Score reflects how strongly the phrase/MeSH term is represented in  $S$  as compared to all of Medline.

We select the top scoring fifty phrases/MeSH terms which are passed to the optimal coverage algorithm described below. While choosing these fifty concepts we limit our attention to the concepts that appear in no more than one hundred documents or half of the size of the set  $S$ , whichever is smaller. This threshold ensures that we identify more specific rather than general concepts.

## Choosing Optimal Coverage

We define the coverage of a concept as the number of documents in  $S$  that contain the concept or its variants. The optimal coverage algorithm is an iterative procedure that starts with the above fifty concepts and identifies among them ten that maximize the total coverage, i.e. the total number of documents containing at least one of these concepts or their variants. By doing so, we diversify the list of concepts that are presented to the user. The resulting ten concepts along with the subsets of documents containing them are presented to the user. These groups of documents are not mutually exclusive as documents are listed with every concept they contain.

## Experiments

We tested our system on five sets describing Cystic Fibrosis, Deafness, Digeorge Syndrome, Autism, and Hypertrophic Cardiomyopathy. Table 2 lists sets of concepts identified for each topic sorted by the negative log of the hypergeometric p-value.

Ten concepts are identified in each of the five sets. These concepts represent a diverse set of topics that naturally subdivide the original set of documents into smaller groups. The results were manually examined and found to be useful.

**Table 1.** A list of text strings recognized as variants of the same ‘attention deficit hyperactivity disorder’ concept.

attention deficit disorder hyperactivity; attention deficit hyperactive disorder; attention deficit hyperactive disorders; *attention deficit hyperactivity disorder*; attention deficit hyperactivity disorders; attentional deficit hyperactivity disorder; hyperactive child syndrome; hyperactive disorder; hyperactive disorders; hyperactivity attention deficit disorder; hyperactivity disorder; hyperactivity disorders; hyperkinetic disorder; hyperkinetic disorders; hyperkinetic syndrome; hyperkinetic syndromes.

**Table 2.** Set of key concepts identified for Cystic Fibrosis (1), Autism (2), Deafness (3), Digeorge Syndrome (4), and Hypertrophic Cardiomyopathy (5).

1: sweat chloride testing; cf family; cystic fibrosis, prevention and control; cystic fibrosis, therapy; cystic fibrosis, psychology; prenatal diagnosis; genetic diseases, inborn; genetic examination; mutation, genetics; gene therapy, methods.

2: rett syndrome; fragile x syndrome, genetics; autistic disorder, epidemiology; developmental disabilities, genetics; asperger's syndrome; transmission disequilibrium; autistic disorder, psychology; chromosomes human, pair 7; attention deficit hyperactivity disorder; genetic predisposition to disease, genetics.

3: non syndromic hearing loss; hearing disorders, genetics; waardenburg's syndrome; deafness, congenital; vestibular aqueducts; hearing loss, sensorineural, physiopathology; induced hearing loss; alport's syndrome; deaf children; hair cells, auditory.

4: cardio facial syndrome; digeorge syndrome, complications; digeorge syndrome, pathology; critical region; digeorge syndrome, diagnosis; abnormalities, multiple, genetics; t-box domain proteins; gene deletion; aortic arch; immunologic deficiency syndromes.

5: cardiomyopathies, genetics; cardiomyopathy, hypertrophic, therapy; cardiomyopathy, hypertrophic, ultrasonography; cardiomyopathy, hypertrophic, complications; myosins, genetics; cardiomyopathy, dilated, genetics; troponine i; cardiomyopathy, hypertrophic, physiopathology; carrier proteins, genetics; mutation, genetics.

## References

W. Kim, L. Yeganova, D. Comeau, W. J. Wilbur, 2012, Identifying well-formed biomedical phrases in MEDLINE text, Journal of Biomedical Informatics.

H. Larson, 1982, Introduction to Probability Theory and Statistical Inference, John Wiley and Sons, New York

## Acknowledgments

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.