# Chapter 7
# Quality-Based Knowledge Discovery from Medical Text on the Web

## Example of Computational Methods in Web Intelligence

Andreas Holzinger, Pinar Yildirim,
Michael Geier, and Klaus-Martin Simonic

**Abstract.** The MEDLINE database (Medical Literature Analysis and Retrieval System Online) contains an enormously increasing volume of biomedical articles. Consequently there is need for techniques which enable the quality-based discovery, the extraction, the integration and the use of hidden knowledge in those articles. Text mining helps to cope with the interpretation of these large volumes of data. Co-occurrence analysis is a technique applied in text mining. Statistical models are used to evaluate the significance of the relationship between entities such as disease names, drug names, and keywords in titles, abstracts or even entire publications. In this paper we present a selection of quality-oriented Web-based tools for analyzing biomedical literature, and specifically discuss PolySearch, FACTA and Kleio. Finally we discuss Pointwise Mutual Information (PMI), which is a measure to discover the strength of a relationship. PMI provides an indication of how more often the query and concept co-occur than expected by change. The results reveal hidden knowledge in articles regarding rheumatic diseases indexed by MEDLINE, thereby exposing relationships that can provide important additional information for medical experts and researchers for medical decision-making and quality-enhancing.

## 1 Introduction

MEDLINE (Medical Literature Analysis and Retrieval System Online) is a bibliographic database for the life sciences and includes bibliographic information for papers of academic journals covering a broad range of biomedical and health care topics. Moreover, MEDLINE covers much of the literature in biology and

G. Pasi et al. (Eds.): *Qual. Issues in the Management of Web Information*, ISRL 50, pp. 145–158.
DOI: 10.1007/978-3-642-37688-7_7

biochemistry. Maintained by the United States National Library of Medicine (NLM), MEDLINE is available for free on the Web and searchable via tools such as PubMed [1] and Entrez [2].The MEDLINE database is the primary resource for biomedical researchers and contains currently 21,763,549 total records [3]. Within this big data, a wealth of scientific information is existing and knowledge on relationships between biomedical concepts including genes, diseases and cellular processes is hidden [4]. All the information contained in the database is stored as text. The rapid growth of these text collections makes it difficult for humans to access the required data in a convenient and effective manner.

Figure 1 shows the vastly increasing number of publications in the MEDLINE database from 1940 until 2011. The number of publications was determined using the PubMed query *"pubyear"[Publication Date]*, where *pubyear* was replaced by the corresponding years. For the year 2011 986,794 publications are listed, in May 2012 already 420,933 publications are found for the year 2012.

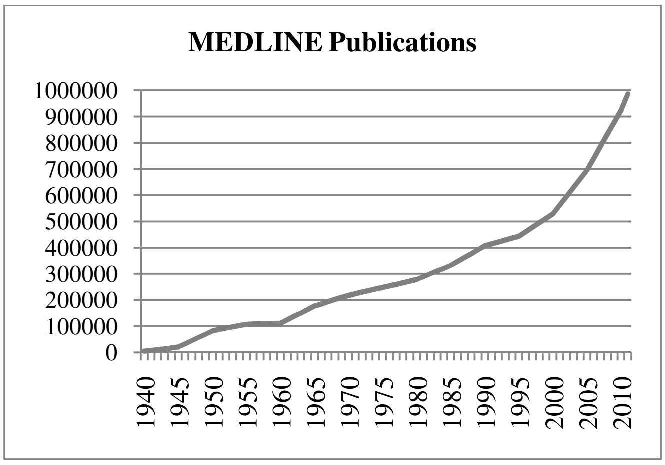


**Fig. 1** Yearly number of MEDLINE publications from 1940 to 2011 (queried in steps of five years on 14/05/2012)

In order to make this data accessible, usable and useful, smart information retrieval systems that can operate on these non-standardized text (often called: “free text”) are essential [5]. Consequently, there is a strong necessity of developing quality-based methods for the extraction of relevant information (such as keywords related with diseases) from the literature, which is written in natural language.

Data mining on text has been designated at various sources as statistical text processing, knowledge discovery in text, intelligent text analysis, or natural language processing, depending on the application and methodology that is used [6], [7].

Text mining aims at developing technologies helping to cope with the interpretation of these large volumes of publications. A commonly used method to establish such relationships between biomedical concepts from literature is co-occurrence analysis. Apart from its use in knowledge retrieval, the co-occurrence method is also well suited to discover new, hidden relationships between biomedical concepts. This technique is applied in text mining and the methodologies and statistical models are used to evaluate the significance of relationship between entities such as disease names, drug names, and keywords in titles, abstracts or even entire publications.

**Table 1** Feature comparison of various biomedical text mining tools [8]

| | **Entrez** | **MedMiner** | **Alibaba** | **PolySearch** |
|---|---|---|---|---|
| **Type of search supported** | Literature, Disease, Gene, Structure, Taxonomy, SNP, Compound, etc. | Gene, Drug, Text Word | Gene, Disease, Drug, Tissues/Organs, Cells, Species | Gene, Disease, Drug, Metabolite, Tissues/Organs, Subcellular Localization, Text Word |
| **Extensive hyperlinking** | Most Extensive | Less Extensive | Less Extensive | More Extensive |
| **Text and sentence highlighting** | No | Yes | Yes | Yes |
| **Co-occurrence scoring scheme** | None | None | Sentence level | Sentence level |
| **Use of keywords for association words** | None | Predefined keywords | Predefined keywords | Predefined & custom association words |
| **Sentence pattern recognition** | No | No | Yes | Yes |
| **Thesaurus query synonym expansion** | Yes, limited | Yes, limited | None | Yes, extensive |
| **Databases** | PubMed, OMIM, Gene, MMDB, Taxonomy, dbSNP, PubChem, etc. | PubMed, GeneCards | PubMed | PubMed, OMIM, Swiss-Prot, DrugBank, HMDB, HPRD, GAD, HapMap, dbSNP, CGAP, HGMD |

## 2 Web-Based Tools for Analyzing Biomedical Literature

There are several Web-based tools for the analysis of biomedical literature. Most of the tools provide the analysis of co-occurrence between biomedical entities such as disease, drugs, genes, proteins and organs. Some provide additional measures, such as Pointwise Mutual Information.

Four tools (Entrez, MedMiner, Alibaba, and PolySearch) are compared in Table 1. [9], however, provides a more extensive overview of Web tools for searching biomedical literature. Kleio and FACTA (see later) are not mentioned, PolySearch, however, and more than 25 tools from the following categories are included:

- Ranking PubMed's search results (example: RefMed)
- Clustering and categorizing results into topics (example: McSyBi)
- Extracting and displaying semantics and relations (example: MEDIE)
- Visualization and improving search interface and retrieval experience (example: iPubMed)

**PolySearch** can produce a list of concepts which are relevant to the user's query by analysing multiple information sources including PubMed, OMIM, Drugbank and Swiss-Prot. It covers many types of biomedical concepts including organs, diseases, genes/proteins, drugs, metabolites, SNPs, pathways and tissues.

The general issue of synonyms and acronyms is handled by PolySearch by optionally automatically expanding the query with synonyms and acronyms. A list of filter words excludes unwanted results. One drawback of PolySearch is the low speed performance of the system.

EBIMed, XplorMed, MedlineR, LitMiner and Anni are commonly used tools and they provide similar functionality with PolySearch [10], [7].

NaCTeM (The National Centre of Text Mining) also develops Web-based tools such as FACTA/FACTA+ and Kleio. These are text search engines for MEDLINE abstracts, which are designed particularly to help users browse biomedical concepts (e.g. genes/proteins, diseases, enzymes and chemical compounds) appearing in the documents retrieved by the query. By revealing associations between biomedical concepts, **FACTA** allows to gain new knowledge from the large amount of MEDLINE text data. The distinct advantage of FACTA is that it delivers real-time responses while being able to accept flexible queries [11]. FACTA covers six categories of biomedical concepts: human genes/proteins, diseases, symptoms, drugs, enzymes and chemical compounds. The concepts appearing in the documents are recognized by dictionary matching. UMLS (Unified Medical Language System) is used for diseases and symptoms. UMLS constitutes a valuable lexical resource integrating a thesaurus and multilingual vocabulary database of health-related concepts as well as the semantic relationships between them. FACTA receives a query from the user as the input. A query can be a concept name like "Rheumatoid Arthritis", a concept ID or a combination of these. The system then retrieves all the documents that match the query from MEDLINE using word/concept indexes. The concepts contained in the documents are then counted and ranked according to their relevance to the query. For the input query "Rheumatoid Arthritis" with disease as a selected concept, and the system retrieves 94834 documents from MEDLINE. The results are displayed as a table and ranked by their frequencies which indicate how many times selected concept appears in the articles with the query word. For example, "Polyarthritis" which is a kind of Rheumatoid disease appears 4393 times with "Rheumatoid Arthritis" [12].

One issue of FACTA is that synonyms and variations of the spelling of terms are often not considered properly. As shown in Figure 3, it is not distinguished between "weakness" and "Weakness", for example.

FACTA+ Visualizer [13] is an Adobe Flash-based browser application which presents the query results of a FACTA query as tile chart (Figure 2, Figure 3).For supporting data analysis by medical experts, who typically are not aware of the mathematical or technical background of text mining tools, good visualisations of the results are essential.

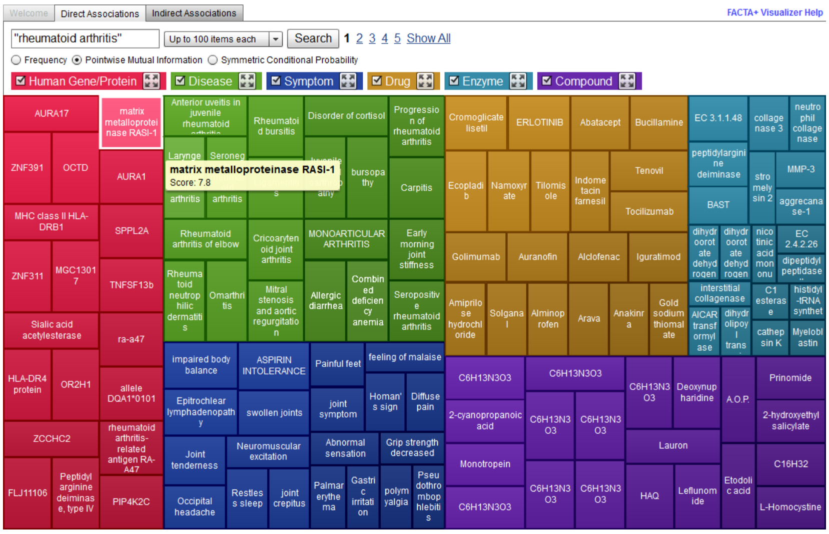


**Fig. 2** FACTA+ Visualizer: Pointwise Mutual Information

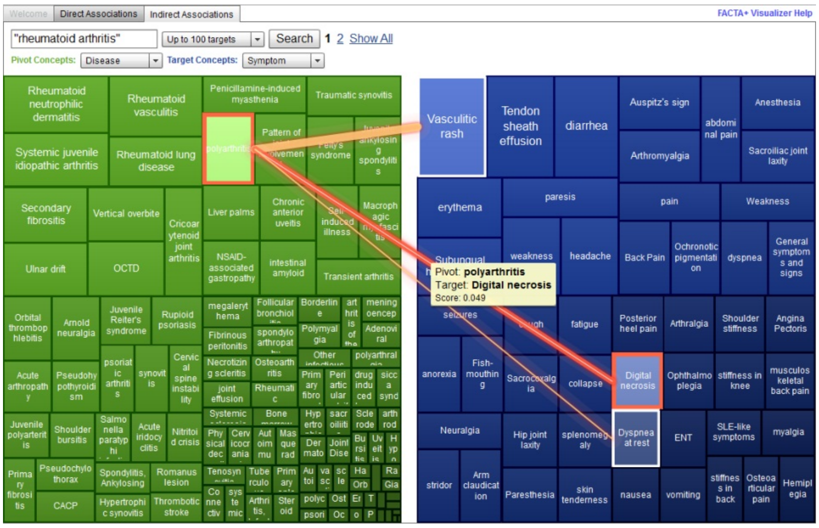


**Fig. 3** FACTA+ Visualizer: Indirect associations between pivot concepts and related target concepts

**Kleio** is an advanced information retrieval (IR) system developed by NaCTeM and offers textual and metadata searches across MEDLINE and provides enhanced searching functionality by leveraging terminology management technologies [14].

Kleio draws upon one of the technologies from the NaCTeM text mining tool kit to enhance automated detection and mark-up of biologically important terms appearing in text, such as gene/protein names. One of these tools is AcroMine, which disambiguates acronyms based upon the context in which they appear. This functionality plays a key role in searching large document collections by allowing users to expand their queries and to include synonymous acronyms without losing the specificity of the original query.

The rich variety of term variants is a stumbling block for information retrieval as these many forms have to be recognized, indexed, linked and mapped from text to existing databases. Typically, most of the currently available information retrieval systems for the biomedical domain fail to deal with the problems of term ambiguity and variability. Kleio addresses this problem for reducing the diversity of term variation. Another key innovation of Kleio is dealing with the variety of names (terms) for denoting the same concept. To map these forms (e.g. IL2, IL-2 and Interlukin-2) to biological databases, machine learning based term normalization techniques which reduce term variation (e.g. il2) is used. An advantage of applying term normalization is to permit efficient look-up and to discover ambiguous and variant terms in the resources [14].

In order to develop a study to discover hidden relationships for biomedical entities such disease-disease relationships, a Web-based text mining tool can be used to find entity names and their co-occurrence frequencies in MEDLINE articles for each entity. Normalisation is another concern for text mining based studies. Biomedical names have some variations such as synonyms. These names need to be normalized to one specific name. For example, Wegener's Granulomatosis and Wegener's Granuloma indicate same diseases and can be mapped to Wegener's Granulomatosis. During the normalisation process, some biomedical resources should be used and interviewing with the biomedical experts can be needed [7].

Statistical techniques also play an important role for text mining studies [15]. There are some measures of co-occurrence analysis. The simplest method to identify relationships is using the co-occurrence assumption: terms that appear in the same texts tend to be related. For example, if a protein is mentioned often in the same abstracts as a disease, it is reasonable to hypothesize that the protein is involved in some aspect of the disease. The degree of co-occurrence can be quantified statistically to rank and eliminate statistically weak co-occurrences.

Web-based tools for discovering such relationships in medical literature may reveal new information and lead to a better understanding of certain concepts and therefore to higher quality of medical treatment.

**Table 2** Comparison of three Web-based tools for analyzing biomedical literature

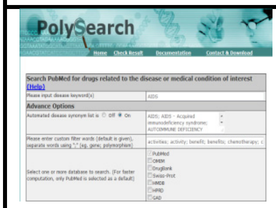
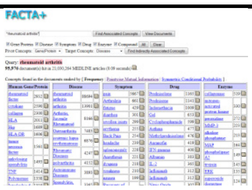
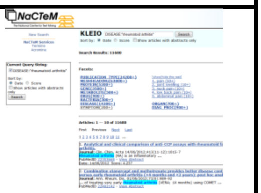
| **PolySearch** | **FACTA+** | **Kleio** |
|---|---|---|
| <br>PolySearch<br>Home Check Result Documentation Contact & Download<br>Search PubMed for drugs related to the disease or medical condition of interest (Help)<br>Please input disease keyword(s): AIDS<br>Advance Options<br>Automated disease synonym list is ○ Off ● On: AIDS; AIDS - Acquired immunodeficiency syndrome; AUTOIMMUNE DEFICIENCY<br>Please enter custom filter words (default is given), separate words using ";" (eg. gene; polymorphism): activities; activity; benefit; benefits; chemotherapy; c<br>Select one or more database to search. (For faster computation, only PubMed is selected as a default): ☑ PubMed ☐ OMIM ☐ DrugBank ☐ Swiss-Prot ☐ HMDB ☐ HPRD ☐ GAD | <br>FACTA+<br>"rheumatoid arthritis"<br>Find Associated Concepts / View Documents<br>Query: rheumatoid arthritis<br>Human Gene/Protein / Disease / Symptom / Drug / Enzyme | <br>NaCTeM<br>KLEIO DISEASE:"rheumatoid arthritis" Search<br>Sort by: ● date ○ score ☐ show articles with abstracts only<br>Facets:<br>1. Analytical and clinical comparison of anti-CCP assays with rheumatoid arthritis.<br>2. Combination etanercept and methotrexate provides better disease control |
| **Scope** | | |
| Finds associated concepts to a given concept. | Finds associated concepts to a given concept. | Search for concepts of certain categories. IR system supported by terminology management technologies |
| **Ranking Algorithm(s)** | | |
| Proprietary PolySearch Relevancy Index, PRI [10] | Co-occurrence Frequency, Pointwise Mutual Information, Symmetric Conditional Probability | Date, Score |
| **Data sources** | | |
| PubMed, OMIM, DrugBank, Swiss-Prot, HMDB, HPRD, GAD, HapMap, dbSNP, CGAP, HGMD | MEDLINE, UniProt, BioThesaurus, UMLS, KEGG, DrugBank | MEDLINE, BioThesaurus, acronym dictionary (mapping created from MEDLINE) |
| **Strengths** | | |
| Use of biomedical thesauruses | Flexible queries.<br>Indexing of concepts (→ quick search results) [11] | Acronym recognition and disambiguation.<br>Normalisation of biology terms.<br>Named entity recognition for gene/protein names.<br>Indexing of terms.<br>Reduction of term variation [14]. |

**Table 3** Comparison of three Web-based tools for analyzing biomedical literature (part two)

| PolySearch | FACTA+ | Kleio |
|---|---|---|
| **Limitations** | | |
| Slow, Finds associated concepts belonging to only one single category. Novel or newly named terms are not recognized (simple dictionary approach to identify biological or biomedical associations) [10] | Limited synonyms/term variation support. | - |
| **Supported concept categories** (in: accepted as input; out: provided as output) | | |
| Disease (in/out), gene/protein (in/out), Drug (in/out), Metabolite (in/out), SNP (RS#) (in/out), Gene sequence (in), Text word (in), Pathway(in/out), Tissue (in/out), Organs (out), Subcellular Localizations (out) | Human Gene/Protein, Disease, Symptom, Drug, Enzyme, Compound | Protein, Gene, Metabolite, Disease, Symptom, Organ, Diagnostic/therapeutic procedure, Medical phenomenon or process, Reagent or diagnostic aid, acronym, author, Publication type |

## 3 Pointwise Mutual Information

A very interesting and useful concept based on information theory is mutual information.

Mutual Information (MI) goes back to Shannon (1948) [16] and is a measure of the mutual dependence between two random variables[1] $X$ and $Y$. The measure itself and the instantiation for specific outcomes are called Pointwise Mutual Information (PMI). It has been introduced to the text mining community by Church & Hanks (1990) [17] as an alternative measure (association ratio) for measuring word association norms, based on the theoretic concept of mutual information. The association ratio can be scaled up to provide robust estimates of word association norms and has up to date proven to be a very useful association measure in Web-based text mining tasks [4].

Mutual Information can be seen as a measure of the information overlap between $X$ and $Y$, where the values have probabilities $p(x)$ and $p(y)$. Consequently, the joint probability of $p(x, y)$ is defined as:

[1] Capitalized variable names refer to random variables.

$$I(X;Y) = \sum_{x,y} p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$

Originally, Fano (1961) [18] used the $\log_2(x)$, however, any logarithm can be used, and changing the base of the logarithm changes the unit of measurement of information [19].

The information overlap between $X$ and $Y = 0$, when the two variables are independent, as $p(x)p(y) = p(x,y)$. When $X$ determines $Y$, $I(X;Y) = H(Y)$, where $H(Y)$ is the entropy of, or lack of information about $Y$, defined as:

$$H(Y) = -\sum_{y} p(y) \log p(y)$$

If $X$ and $Y$ are perfectly correlated, i.e. they determine each other, then $I(X;Y)$ reaches a maximum $H(X) = H(Y) = H(X,Y)$, where $H(X,Y)$ is the joint entropy of $X$ and $Y$.

This leads to the definition of Fano (1961), who stated, that if two points $P$ (information objects, e.g. words), $x$ and $y$, have probabilities $P(x)$ and $P(y)$, then their Pointwise Mutual Information, $PMI(x,y)$ is defined as:

$$PMI(x,y) = \log\left(\frac{P(x,y)}{P(x)P(y)}\right)$$

In a recent study on disease-disease relationships for rheumatic diseases by Yildirim, Simonic & Holzinger (2012) this measure was used to discover the strength of a relationship and to provide an indication of how more often the query and the concept co-occur. After ranking of the measures and the frequencies together, the results revealed hidden knowledge in articles regarding rheumatic diseases indexed by MEDLINE. Such relationships can provide important additional information for medical experts and researchers for medical decision-making [4]. In its original form, the method is restricted to the analysis of two-way co-occurrences. Problems involving natural language processing, however, need not to be restricted to two-way co-occurrences; often, a particular problem can be more naturally tackled if it is formulated as a multi-way problem; consequently the framework of tensor decomposition, that has recently been introduced analyzes language issues as multi-way co-occurrences [20].

It was shown by [21] that a version of PMI trained on Wikipedia outperformed several publicly available measures of semantic relatedness and might even outperform LSA (latent semantic analysis) when trained with a sufficiently large amount of data. For practical applications this is interesting because similarity judgments are fast and easy to calculate using PMI, even on huge data sets. Furthermore, [22] showed that PMI based topic models coincide well with human perception. It can be summarized that the previously mentioned sources show the eligibility of PMI for different scenarios in knowledge discovery tasks.

Pointwise Mutual Information is an ideal measure of word association norms based on information theory and we selected this measure to analyze rheumatic diseases. PMI compares the probability of observing two items together with the probabilities of observing two items independently. Therefore, it can be used to estimate whether the two items have a genuine association or are observed at random [12].

Let two words, $w_i$ and $w_j$, have probabilities $P(w_i)$ and $P(w_j)$. Their mutual information $PMI(w_i, w_j)$ is defined as:

$$PMI(w_i, w_j) = \log\left(\frac{P(w_i, w_j)}{P(w_i)\,P(w_j)}\right)$$

The other way to discover hidden knowledge between biomedical entities is to use machine learning techniques to analyse the articles. At first, the co-occurrence frequencies of entity-entity can be extracted by using biomedical text mining tool. Most common of them are selected and normalized for each entity to create datasets. For example, relationships between diseases and symptoms can be explored. The frequencies of symptoms for each disease are found by using Web-based biomedical text mining tool. The frequency of the symptom provides the number of times a considered symptom appears in the articles. After normalizing the names, the dataset containing diseases and the frequencies of symptoms can be created. At the last stage, machine learning algorithms can be applied on the dataset to discover similarity between diseases. For instance, cluster analysis can be used to analyse the dataset. Cluster analysis is one area of machine learning of particular interest to data mining.

A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Cluster analysis has been also widely used in numerous applications, including pattern recognition, data analysis, image processing and biomedical research. In biomedical text mining studies, cluster analysis can be used to explore similarities between entities such as diseases, gene and organs [23].

## 4 Symmetric Conditional Probability

FACTA+ not only allows calculating co-occurrence frequencies and PMI, but also the symmetric conditional probabilities (SCP) for identifying associated concepts. [24] proposed SCP as measure for testing the correlation between terms x and y by multiplying the conditional probabilities of x given y and y given x.

$$SCP(x,y) = p(x|y) \cdot p(y|x) =$$
$$\frac{p(x,y)}{p(y)} \cdot \frac{p(x,y)}{p(x)} = \frac{p(x,y)^2}{p(x) \cdot p(y)}$$

## 5 FACTAs Scoring Methods: Frequency, PMI, and SCP

In the following paragraphs we will exemplarily compare three scoring methods used by FACTA to rank the associated concepts to a given concept, specified by a textual query.

Table 4 shows the first 27 of 379 results for the query “rheumatoid arthritis” using FACTA+. The related concepts to the search term are listed in descending order, ordered by the score describing the relation to rheumatoid arthritis. Three scoring methods were compared: Frequency of co-occurrence (Frequency), Pointwise Mutual Information (PMI), and Symmetric Conditional Probability (SCP).

In order to get an impression of the “agreement” amongst the three methods, Kendall's coefficient of concordance (Kendall’s W) was calculated [25]. Kendall’s W describes the agreement amongst raters concerning the ranking of items. In this case the ranking of the retrieved associated concepts is determined by the strength of the relation to rheumatoid arthritis. The “raters” are the tested methods Frequency, PMI, and SCP. A Kendall’s W value of 1 means complete agreement amongst the raters, a value of 0 means no agreement. Kendall’s W for all three methods the overall agreement (the agreement over all 379 result items) is 0.3778. Looking only at Frequency and PMI, the value is 0.5214. For PMI and SCP the value is 0.5577, and for Frequency and SCP it is 0.5210. We can see that, when looking at no more than two methods at the same time, PMI and SCP have a slightly higher agreement than the other combinations of two methods.

However, when looking at the 27 highest rated terms of each method (see Table 4) it can be observed that the methods Frequency and SCP top-rank similar terms while PMI top-ranks different terms.

In practice, when medical professionals use such tools to discover new relations between concepts, especially the highest ranked results are of importance. As mentioned before, in several studies it was shown that PMI has high performance for certain scenarios by not being computationally intensive at the same time. This makes PMI a good candidate for creating high quality Web-based applications.

**Table 4** Comparison of FACTAs ranking of related concepts from the category Symptom for the query “rheumatoid arthritis” created by the methods co-occurrence frequency, PMI, and SCP

| **Frequency** | | **PMI** | | **SCP** | |
|---|---|---|---|---|---|
| pain | 5667 | impaired body balance | 7,8 | swollen joints | 0.002 |
| Arthralgia | 661 | ASPIRIN INTOLERANCE | 7,8 | pain | 0.001 |
| fatigue | 429 | Epitrochlear lymphadenopathy | 7,8 | Arthralgia | 0.001 |
| diarrhea | 301 | swollen joints | 7,4 | fatigue | 0.000 |
| swollen joints | 299 | Joint tenderness | 7 | erythema | 0.000 |
| erythema | 255 | Occipital headache | 6,2 | splenomegaly | 0.000 |
| Back Pain | 254 | Neuromuscular excitation | 6,2 | Back Pain | 0.000 |
| headache | 239 | Restless sleep | 5,8 | polymyalgia | 0.000 |
| splenomegaly | 228 | joint crepitus | 5,7 | joint stiffness | 0.000 |
| Anesthesia | 221 | joint symptom | 5,5 | Joint tenderness | 0.000 |
| dyspnea | 218 | Painful feet | 5,5 | hip pain | 0.000 |
| weakness | 210 | feeling of malaise | 5,5 | metatarsalgia | 0.000 |
| nausea | 199 | Homan's sign | 5,4 | Skin Manifestations | 0.000 |
| Recovery of Function | 193 | Diffuse pain | 5,2 | neck pain | 0.000 |
| low back pain | 167 | Palmar erythema | 5,2 | Eye Manifestations | 0.000 |
| abdominal pain | 141 | Abnormal sensation | 5,2 | low back pain | 0.000 |
| cough | 126 | Gastric irritation | 4,8 | dyspnea | 0.000 |
| analgesia | 120 | Grip strength decreased | 4,8 | weakness | 0.000 |
| Pain, Postoperative | 112 | polymyalgia | 4,8 | Fever of Unknown Origin | 0.000 |
| vomiting | 106 | Pseudothrombophlebitis | 4,7 | nausea | 0.000 |
| neck pain | 105 | Deep granuloma annulare | 4,6 | dry eye | 0.000 |
| collapse | 103 | Axillary lymphadenopathy | 4,5 | diarrhea | 0.000 |
| discomfort | 101 | Calf pain | 4,5 | Epitrochlear lymphadenopathy | 0.000 |
| discomfort | 97 | gastrointestinal colic | 4,3 | ASPIRIN INTOLERANCE | 0.000 |
| Fever of Unknown Origin | 81 | Radiating pain | 4,3 | impaired body balance | 0.000 |
| myalgia | 79 | Musculoskeletal symptoms | 4,3 | Recovery of Function | 0.000 |
| Eye Manifestations | 78 | Arthralgia | 4,2 | myalgia | 0.000 |

# 6 Conclusion

Optimal tools for quality-based text mining and knowledge discovery are of high importance for the MEDLINE database as it is growing extremely fast and will possibly grow even faster in the future. Without such tools many publications will not be noticed by biomedical professionals, consequently much potentially useful information may be lost. Additionally, yet hidden knowledge can be made visible with knowledge discovery tools.

There is a large amount of Web-based tools available which make it possible to search the MEDLINE database and which allow the discovery of new knowledge, such as hidden relations between concepts. In this work we discussed and compared PolySearch, FACTA, and Kleio, while at the same time had a look on FACTAs ranking algorithms for associated concepts and Pointwise Mutual Information (PMI).

The quality of the results and therefore the applicability and the relevance of the algorithms used for text mining are essential. Moreover, the user interface of Web-based tools must not be neglected to support the accessibility medical professionals in an intuitive and effective way.

## 7 Future Work

A large number of Web-based tools are available for searching MEDLINE and for supporting knowledge discovery from the MEDLINE data. However, there are still many issues for research in this area: At first, the non-standardized nature of text is still a big issue, and there is much work left for improvement in the area of synonym recognition as, for example, could be seen during our investigation. Second, it can be stated that for efficient performance, the response time of the Web-based tool must be optimized. Therefore, further investigation is necessary in the optimization of existing algorithms as well as in optimal usage of the available server infrastructure in order to deliver results as quickly as possible. Third, research in end-user centred visualisation and visual analytics of the results is urgently needed in order to support efficient and fast sensemaking processes amongst medical professionals.

## References

1. http://www.ncbi.nlm.nih.gov/pubmed
2. http://www.ncbi.nlm.nih.gov/Entrez
3. http://www.nlm.nih.gov/bsd/revup/revup_pub.html#med_update
4. Holzinger, A., Simonic, K.M., Yildirim, P.: Disease-disease relationships for rheumatic diseases. In: COMPSAC 2012, Izmir, Turkey (2012) (in print)
5. Kreuzthaler, M., Bloice, M.D., Faulstich, L., Simonic, K.M., Holzinger, A.: A Comparison of Different Retrieval Strategies Working on Medical Free Texts. Journal of Universal Computer Science 17, 1109–1133 (2011)
6. Solka, J.L.: Text data mining: theory and methods. Statistics Surveys 2, 94–112 (2008)
7. Yıldırım, P., Çeken, Ç., Çeken, K., Tolun, M.R.: Clustering Analysis for Vasculitic Diseases. In: Zavoral, F., Yaghob, J., Pichappan, P., El-Qawasmeh, E. (eds.) NDT 2010. CCIS, vol. 88, pp. 36–45. Springer, Heidelberg (2010)

8. http://wishart.biology.ualberta.ca/polysearch/cgi-bin/help.cgi#eval1
9. Lu, Z.: PubMed and beyond: a survey of web tools for searching biomedical literature. Database 2011 (2011)
10. Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S., Wishart, D.S.: PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. Nucleic Acids Research 36, W399–W405 (2008)
11. Tsuruoka, Y., Tsujii, J., Ananiadou, S.: FACTA: a text search engine for finding associated biomedical concepts. Bioinformatics 24, 2559–2560 (2008)
12. Yildirim, P., Çeken, Ç., Hassanpour, R., Tolun, M.R.: Prediction of similarities among rheumatic diseases. Journal of Medical Systems, 1–6 (2010)
13. http://refine1-nactem.mc.man.ac.uk/facta-visualizer/
14. Nobata, C., Cotter, P., Okazaki, N., Rea, B., Sasaki, Y., Tsuruoka, Y., Tsujii, J., Ananiadou, S.: Kleio: a knowledge-enriched information retrieval system for biology (Year)
15. Schmeier, S., Hakenberg, J., Kowald, A., Klipp, E., Leser, U.: Text mining for systems biology using statistical learning methods, pp. 125–129 (Year)
16. Shannon, C.E.: A Mathematical Theory of Communication. Bell System Technical Journal 27, 379–423 (1948)
17. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics 16, 22–29 (1990)
18. Fano, R.: Transmission of Information: A Statistical Theory of Communications. MIT Press, Cambridge (1961)
19. Bouma, G.: Normalized (pointwise) mutual information in collocation extraction. From Form to Meaning: Processing Texts Automaticallym. In: Proceedings of the Biennial GSCL Conference, pp. 31–40. Günter Narr Verlag, Tübingen (2009)
20. Van de Cruys, T.: Two multivariate generalizations of pointwise mutual information. In: Workshop on Distributional Semantics and Compositionality (DiSCo 2011), pp. 16–20. Association for Computational Linguistics (Year)
21. Recchia, G., Jones, M.N.: More data trumps smarter algorithms: comparing pointwise mutual information with latent semantic analysis. Behavior Research Methods 41, 647–656 (2009)
22. Newman, D., Noh, Y., Talley, E., Karimi, S., Baldwin, T.: Evaluating topic models for digital libraries. In: Proceedings of the 10th Annual Joint Conference on Digital Libraries, pp. 215–224. ACM, Gold Coast (2010)
23. Takada, T.: Mining local and tail dependence structures based on pointwise mutual information. Data Min. Knowl. Discov. 24, 78–102 (2012)
24. Ferreira da Silva, J., Pereira Lopes, G.: A local maxima method and a fair dispersion normalization for extracting multiword units from corpora. In: Sixth Meeting on Mathematics of Language, pp. 369–381 (Year)
25. Bar-Ilan, J.: Comparing rankings of search results on the web. Inf. Process. Manage. 41, 1511–1519 (2005)