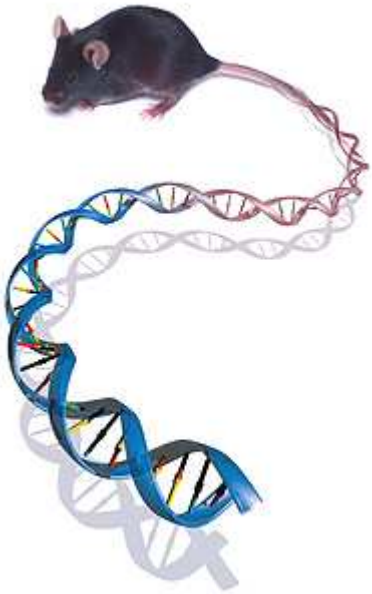
Universiteit Twente
faculteit der informatica

Graduation report:
Managing Genome databases



An attempt with "out of the box" database components.

By:
Bart van Borssum Waalkes (s9705597)

Supervision:
dr.ir. D. Hiemstra
dr. P.E. van der Vet

Utrecht, 2005

Abstract

Over the last 20 years genomics research has gained a lot of interest. Every year millions of articles are published and stored in databases. Researchers around the world want to be able to search for information about e.g. genes, diseases and enzymes. As of this moment there are no search methods available that give researchers a viable and efficient way to search for information about genomics data.

This Report discusses how information can be found using a desktop pc and a widely available database system. It will describe how the documents are found as well as the precision and recall of a query.

With the help of several well know Information retrieval methods, such as Boolean retrieval, TF*IDF and stemming, the effects of these searching methods will be tested, and compared to each other.

The effects these methods have on the overall results, of the system, will be evaluated and the system will be compared to other systems what are using the same documents and questions. After all the results have been evaluated a few hints will be given for ways to improve the system.

Preface

This paper is written as a conclusion to my studies at the University of Twente.

The genomics TREC was used as a basis for my gradation assignment. The Medline data and the questions they provided were used as a benchmark for the developed system.

I would like to thank dr.ir. D. Hiemstra and dr. P.E. van der Vet for their help and supervision during this project.

Utrecht, 2005

B.G. van Borssum Waalkes

Table of contents

1 Introduction

The Text REtrieval Conference (TREC) supports research within the information retrieval community. TREC has several branches (tracks) of research going every year. One of these tracks is the "***Genomics Track***".

The Genomics track was introduced in 2003 and will run for at least 5 years. The purpose of the track is to study information retrieval in the genomics data domain (this includes not just gene sequences, but also documentation such as research papers, lab reports, etc).

For my graduation assignment I joined the 2004 genomics track, more specifically the "Ad Hoc Retrieval Task".

> The structure of this task is a conventional searching task based on a 10-year subset of MEDLINE (about 4.5 million documents and 9 gigabytes in size) and 50 topics derived from information needs obtained via interviews of real biomedical researchers. [008]

Over the last twenty years genomics research has expanded greatly. Every year thousands or articles, papers and journals are published. Most of these articles, papers and journals are included in the MEDLINE database. Despite the fact that almost all information is available, in electronic form; most people still rely on MEDLINE when searching for information. Because MEDLINE is the number one source of information, when searching for genomics information, the ability to search the MEDLINE database is getting more and more important. The Genomics track tries to find new and more efficient ways to search the genomics information.

Before we go into the goal of this paper, a closer look will be taken at the other participants of TREC. An overview will be given of the problems and expectations other researchers are having.

When examining the papers the other participants of the Genomics TREC submitted then there are three problems when searching the TREC database.

- Linguistic problems: synonymy and Homonymy
  Researchers all over the world are using different terms for the same genes, conditions and/or effects.
- Automated query generation.
  The topics provided by TREC are writing in natural language. How can the system atomically generate the query needed to find relevant information out of the topic?
- Huge database

Because of the thousands of articles, papers and journals that are being published the amount of information that has to be searched is huge.

Most of the participants of the Genomics TREC agree that synonymy is the biggest problem when searching for information. They don't however agree on the way to overcome the problem. The solutions to the problem can however be categorized into 2 distinct categories.

- Use of a thesaurus.
  Some participants use a precompiled thesaurus to match the different synonyms with each other [018] and [019].
- Relevance feedback.
  Some participants look at the make up of articles that get a high ranking and expand they search terms with words that are likely to be synonyms of terms they are looking for [012] and [017].

The second problem that people are facing is the automated query generation. TREC gives its member the option to turn in both manual and automatic runs. Manual runs are runs where participants select their own query terms based on their interpretation of the natural langue and the terms they think are important to find information. On automatic runs the system itself will determine which terms it will use to find relevant information. There are several ways of generating automatic queries; a few of them are listed below.

- Using term frequency
  Words in the topics that don't occur at a high frequency in the database are more likely to be of value then words that occur frequently in the database [012] and [017].
- Using controlled vocabularies
  Words that occur in medical dictionaries have a higher likeliness of being relevant. Also words that do not occur in a regular dictionary (proper English) have a high likely hood of being relevant [018] and [020].

After evaluating the problems the other participants identified there are two problems we will be focusing on in this paper.

- A huge dataset:
  For this assignment 9 Gb of raw text needs to be searched.
- No clear naming schema for gene-names
  Synonyms are a big problem when looking for information.

The problem of automated query generation will not be covert because we will use manual runs. The problem of Homonymy will not be looked at. This is because it is a far less serious problem then synonymy because this problem

corrects itself when more search terms are added (the chance that a document contains a homonym of the term you are actually looking for diminishes with every term added).

The goal of this research is to create a search mechanism that researchers can use to find accurate and relevant data about genomics data, which responds within a reasonable time (seconds rather then minutes). We will attempt to achieve this goal using widely available database tools and components. The system runs on a regular desktop computer and all software used can be freely downloaded from the internet. In order to get the best possible result several well known search and/or ranking methods were used to compare their effectiveness.

2 Various search methods

In this chapter I will discuss several popular ways to retrieve data from a dataset. This chapter is divided in 2 sections. The first section will discuss several Retrieval methods. The second section will discuss several retrieval utilities which can be used with any retrieval method to improve the results.

2.1 Retrieval methods

In this section two retrieval methods are discussed. There are many more ways to retrieve data. I chose to only include to most commonly used methods. The methods that aren't listed would either take too much time to implement or wouldn't run efficiently on a desktop system.

Generally retrieval methods can be divided into two classes:

- Exact-match searching. This method only retrieves documents that match all the terms the user is looking for. Examples of exact match searching are Boolean search and Boolean TF*IDF search.
- Partial-match searching. This method retrieves all documents that match at least one of the search terms. The documents are then ranked by the system to ensure that documents that match the search criteria "better" are given a higher rank. Examples of partial match searching are fuzzy set, vector space, and probabilistic retrieval.

Because the search terms for the fifty topics, which TREC provided, are handpicked by the user, all the terms, which are searched for, are of great importance. This means that documents that do not satisfy at least one of the terms have a big chance of being irrelevant. This means the system will need Exact-match searching. In the following section two exact-match search methods are discussed.

2.1.1 Boolean retrieval

Boolean retrieval is the most simple of all retrieval methods. It's called Boolean because the terms of the query are linked together using *AND, OR* and *NOT.* The Boolean retrieval method only retrieves documents that match the query exactly and doesn't have any way of ranking the documents for relevance.

2.1.2 Boolean retrieval with TF*IDF

This method is actually an extension of Boolean retrieval. It was designed in order create a way to rank documents on relevance. The method assigns every term in the database a weight that can be used to judge the importance of that term (terms that occur few times in the database get a high relevance, common terms get a low relevance).

The method works like this:

- A list of documents is retrieved using the "normal" Boolean search.
- For every document
  - For every Term
    - Calculate the TF*IDF ranking of the term in the retrieved document
  - Add all the TF*IDF rankings. This number is the relevance judgment for the document
- Order all the documents on their relevance value.

TF stands for "*Term frequency"* and IDF stands for "*Inverse document frequency"*. The formulas to calculate both TF and IDF are given below.

TF(term) = the frequency of a term in a document.

IDF (term, document) = $\log \frac{\text{documents in the database}}{\text{documents with the term}} + 1$

IDF (term, document) = 0
If no documents are retrieved

The weight of a term in a document can be computed now.

Weight (term, document) = TF(term, document) * IDF(term)

The TF*IDF ranking discussed in this section is the most basic form of TF*IDF rankings. The ranking can be modified to use document specific information (e.g. document length) to further increase the precision of the ranking.
More information about TF*IDF can be found in [002]

2.2 Retrieval utilities

In this section several retrieval utilities are discussed that can be used to improve results of the retrieval methods discussed in the previous section.

2.2.1 Clustering

Clustering tries to group documents by content. This reduces the search space needed by a query to respond. The biggest problem with clustering is the computational complexity. In order to assign every document a cluster the system needs to compare every document with all the other documents. Over the years a lot of methods were created to automate the clustering process. A detailed review of clustering algorithms is given in [013].

2.2.2 Parsing

Using parsing on the documents before they are being searched can improve performance and precision of the system. There are 3 ways to change the documents before they are added into the system:

- Removal of punctuation and case folding.
  - This step usually improves both precision and performance of the system because it lowers the amount of terms available in the system and it decreases the chance of a word getting indexed twice (e.g. Iron-Regulator and ironregulator will both be indexed as ironregulator).
- Removal of stop words.
  - Stop words are words that are used often in sentences but they don't say very much about the meaning of the sentence. Removing these words can greatly decrease the amount of data stored in the database. But special care has to be taken to make sure no search terms occur in the stop words list.
- Stemming of the words.
  - Stemming is a technique for reducing words to their grammatical roots. There is much discussion about the use of stemming in databases. It improves performance of system but it doesn't always improve precision.
  - The popular stemming algorithms are:
    - Lovins stemmer [014].
    - Porter stemmer [007].

2.2.3 Thesauri

The definition of a Thesaurus:

> *A Thesaurus is a type of dictionary where groups of words with the same meaning are grouped together. [015]*

When searching through documents it is often useful for the user to also find synonyms of the words he is looking for e.g. (car and automobile). A thesaurus can help a database system to expand a query so it also includes synonyms, of the search terms the user is looking for.

2.2.4 Multi-word terms

Multi-word terms are search terms that consist of more then one word (e.g. Iron-regulated transporter 1). Using a standard Boolean search the system would search for the separate words of the term and take the intersection of the results of all the separate words. When adding support for multi-word terms, the system will not just take the intersection of the separate words, but will also look at the proximity of the words. If the separate words of the query are not next to each other then the term as a whole will not be counted as a hit.

2.2.5 Mesh terms

A commonly used tool for searching databases is the use of mesh terms. Mesh terms are terms that are added to a document by the administrator of the database. Mesh terms give an abbreviation of the contents of the document. In general the amount of mesh terms defined for a document is lower then the amount of text that would have to be searched otherwise. While mesh terms are a good way to create a fast way to search though documents it is also easy to miss documents. It is impossible to include every subject of a document in the mesh terms (defining to many mesh terms lowers the effectiveness of them).

3 Research directions

Now that the methods of searching though text documents are clear, we will look at the data and the questions that have to be answered.

3.1 The data

The data that will be used is provided as a 9 Gb text file. Every line of the file corresponds with a field from the MEDLINE database. A detailed list of possible fields in the text file can be found in appendix I. Out of all the fields it was decided to only use the abstract, title and mesh terms of every document. The other fields don't give information about the content; they focus primarily on the authors, copyrights and the date and/or place of publishing.

In order to test reliability of the fields that will be used in the tests 10.000 random documents were selected and checked for their title, abstract and mesh term contents. Of all the documents in the test collection over 99% of them have a title defined, 65% has an abstract defined and about 80% has mesh terms defined

The tile of most documents is about 2 lines of text; the abstract has an average of 10 lines of text. In total 4.5 million documents will be searched.

3.2 The topics

TREC provided its participants with a list of 50 topics that need to be answered. For every topic 4 fields are specified.

- ID - This is the number of the topic.
- Title - This is the title of the topic
- Need - This field describes what the user is supposed to search for
- Context - This gives background information why the information is needed.

The list of topics can be found in appendix II.

After examination of the topics they can be divided into 3 groups.

- Type 1
  - topics of the form: find information about A in B (topics: 1, 3, 4 , 6, 9, 10, 11, 13, 14, 15, 18, 20, 21, 23, 26, 27, 31, 32, 34, 38, 39, 40, 44, 46, 47, 48, 50)
- Type2
  - Topics of the form: find a protocol, method or function that describes C (topics: 2, 5, 12, 16, 17, 22, 24, 25, 29, 30, 33, 35, 36, 37, 41, 42, 43, 45, 49)
- Type3
  - Topics of the form: find correlation between D and E (topics: 7, 8, 19, 28)

Where

- A is a protein, gene, disease or process
- B is an organism or a process
- C is a protein, gene, disease or process
- D is a protein, gene, disease or process
- E is a protein, gene, disease or process

3.3 The method

After looking at the data and the topics I've decided to use Boolean searching for the database. This was done because of the need to search for just a few terms per topic and all terms should be present in the document in order for the document to be relevant. Other searching methods (like probabilistic or vector based searching) could also be used but the extra resources needed for those systems are simply not needed when searching terms of equal importance.

For topics of type 1 the A and B need to be present, for topics of type 2 condition C needs to be present and for topics of type 3 both D and E need to be present. If the conditions mentioned for every type of topic do not occur in the fields

specified in section 3.2, then the chances of the document being relevant are very slim.

Once Boolean searching was set it was decided to use the PostgreSQL database system [009]. This system was chosen for several reasons.

- It is widely available for download on the internet.
- It has support to run on Microsoft Windows, the operating system that is running on the desktop pc used to create the system.
- Support for big tables (a table limit of 16 Terabyte).
- Proven stability and performance; postgreSQL has been around since 1987 and has proven itself during the years of its development.
- Excellent support to create stored procedures.

To optimize performance and precision the following utilities will be used:

- TF*IDF
- Parsing.
  - Stemming.
  - Removal of stop words.
  - Removal of punctuation.
- Thesaurus
- Clustering
- Mesh terms

More details on the use of these utilities will be given in the next chapter.

4 Approach

In order to test the effect of the various searching utilities several test runs were conducted. First a database was created that only used the removal of punctuation and clustering. The results we got from this database were used to test the effect of adding several Utilities. New utilities were added to the system every run. After every run the results of that run were compared to the best run up to that moment, if the results had improved the utility was used in all the following runs. If the results had decreased the utility was dropped.

At first only the title and abstract fields of the documents were searched. Once the optimal searching methods for those fields were decided, the effect of adding a search of the mesh terms was added. This was done to limit the amount of data the system had to search.

4.1 Thesaurus

The first addition to the system was a thesaurus. The thesaurus was manually filled with data. The data was retrieved from the “Entrez gene” database [006], and contains synonyms and abbreviations of terms used to search the database. Of the fifty topics provided by TREC only ten contain terms that can be expanded by looking at the “Entrez gene” database. Therefore the effect the thesaurus had on the results was limited to those ten topics.

4.2 Removal of stop words

The second addition to the system was the removal of stop words. The list of stop words I choose to use is the same list of stop words that’s used by the MEDLINE database. The list was created by selecting the words that occur most in the database. Because of the huge amount of hits these words would generate they are useless to use in a query. To be certain, no important words were dropped. The list of fifty topics was compared with the stop words and none of the topics included a search term that was included in the list of stop words. A list of the words that will be removed can be found in Appendix II.

4.3 Stemming

The third addition to the system was stemming. The system uses the widely available Porter stemmer [007]. The porter stemmer attempts to rewrite all words in the database to their stem. This way the chance that a word is missed, because it is not in its base form, is reduced.

4.4 Ranking

Two types or ranking were used. In the following sections both will be discussed. During the first test the effect of using ranking of type 1 was tested.

4.4.1 Ranking type 1

This is a very basic ranking system; all terms that are found have the same importance. The system calculates how often the terms the user is looking for are used in both the title and the abstract and awards 3 point to a hit in the title and 2 points to a hit in the abstract.

4.4.2 Ranking type 2

During the third test a modified TF*IDF system was added. The system does not look at individual terms when creating the weight of a term, but looks at all synonyms of a word as if they are one word. The modified system works like this:

TF(term) = the frequency of a term in a document.
$TF(term)_{total}$ = sum of all TF(term) where the terms are synonyms

$$IDF(term, document)_{total} = \log \frac{\text{documents in the database}}{\text{documents with term or its synonyms}} + 1$$

$IDF(term, document)_{total} = 0$
if no document and synonyms are found.

The weight of a term and its synonyms in a document can be computed now.
$Weight(term, document)_{total} = TF(term, document)_{total} * IDF(term)_{total}$

It was decided to use the most basic form of TF*IDF ranking, this is because it minimizes the computational complexity but still gives a good comparison between the two ranking types. If eventually we want to optimize the results then different ways of expanding the TF*IDF system can be considered.

4.5 Multi-word expressions

The fourth and final change that was added was support for finding multi-word-expressions. This addition was used to determine if names consisting of several words (e.g. Iron-regulated transporter 1) actually occur as one term in the document.

5 Implementation

In this chapter an explanation will be given on how the utilities discussed in the previous chapter can be implemented. In the first section the basic system without any of the features will be discussed. The sections following 5.1 will discuss how the utilities discussed in the previous chapter will be implemented. For every utility added it is assumed that the utilities discussed in the sections before the one being discussed are all implemented.

5.1 The basic system

The system is depicted in figure 5.1. As can be seen the system contains four processes.

- Pre-processing
- Indexing
- Searching
- Ranking

The first two steps, *pre-processing* and *indexing,* are only preformed once, when the database is being initialized. After its initialization the database will be ready to be used. In this chapter all the steps needed to create the database and to get results from it will be discussed. All these steps have to be taken for both the title and the abstract of a document. In the next sections the four processes are explained.

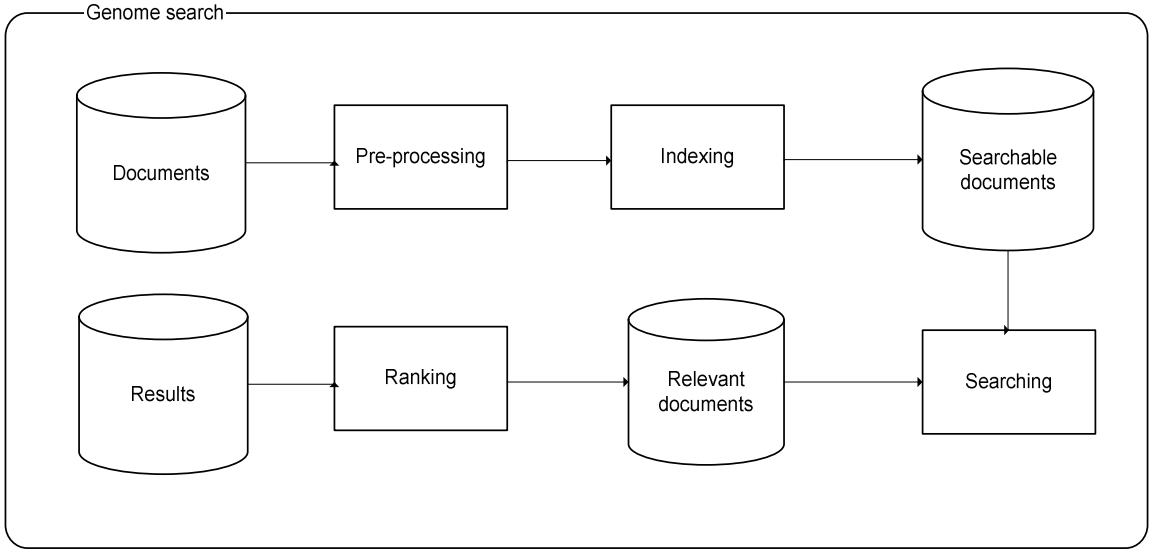


Figure 5.1 – overview of the basic search system

5.1.1 Pre-processing

During this step all characters will be converted to lowercase and all punctuation signs and stopwords will be deleted. The removal of punctuation signs will follow the rules as they are set in Appendix IV.

5.1.2 Indexing

During this process all the documents are put into the database. Figure 5.2 shows the database schema of the database. The indexing process both indexes the abstract and the title fields of the dataset. Both fields are stored in a separate table. The indexing process will give every word it encounters, in the dataset, a unique id. Then for every document a list with the document id and the word id is stored in either the abstract or title table.

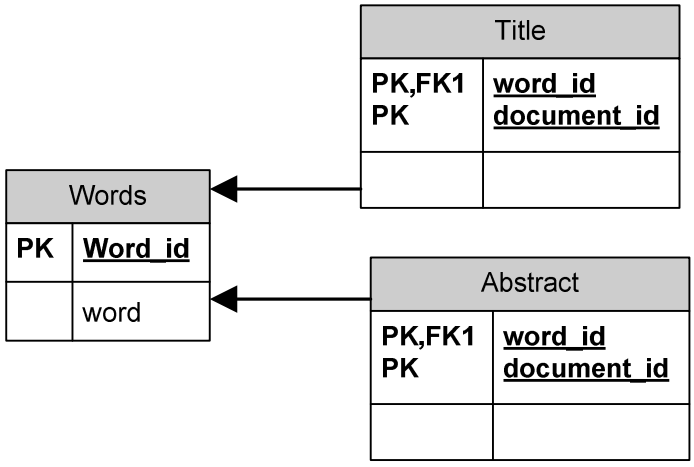


Figure 5.2 – Database schema of the database

5.1.3 Searching

During the searching process the system attempts to retrieve as many relevant documents as possible. It attempts to do this in the following way. At this point the system does not support multi-word terms, the words of multi-word terms are treated as separate terms.

- System receives a query with X terms
- For every X retrieve the word id(W) of X
- For every X retrieve the document ids from both the title and abstract tables of the documents that contain W.
- Take the Intersection of all the documents retrieved for every X.

At the end of this process the system will have a list of all documents containing all the terms.

5.1.4 Ranking

During the ranking process documents are listed in order of importance. The basic system does not have any sort of ranking, so during the ranking process the data will not be changed. This process was included in figure 5.1 because it will be needed in future improvements discussed in the following sections.

5.2 Thesaurus

In order to make use of a thesaurus the database schema (figure 5.2) has to be changed. The new database schema is given in figure 5.3. The thesaurus is manually generated after the system is about to finish the indexing process.

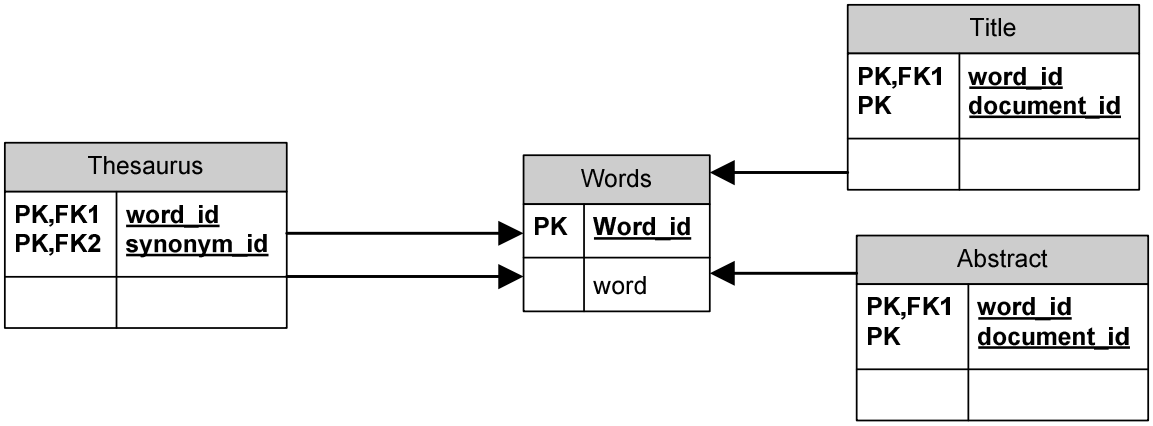


Figure 5.3 – Database schema including a thesaurus

As can be seen in figure 5.3 the thesaurus contains a list of word ids. It is basically linking words together if they are synonyms.

5.2.1 Searching with the thesaurus

During the searching process the system attempts to retrieve as many relevant documents as possible. It attempts to do this in the following way.

- System receives a query with X terms
- For every X retrieve the word id(W) of X
- For every W retrieve synonyms Y from the thesaurus
- For every X retrieve the document ids from both the title- and abstract tables of the documents that contain W or Y.
- Take the Intersection of all the documents retrieved for every X.

At the end of this process the system will have a list of all documents containing all the terms and/or their synonyms.

5.3 Stemming

Like the removal of stop words, stemming will be added in the pre-processing stage, but it will be put between the removal of punctuation and the removal of stop words. This is because stemming might result in more stop words appearing in the document. These stop words then have to be removed.
Stemming has no influence on the design of the database. The database remains as depicted in figure 5.3. Stemming does have a small impact on the searching process. Therefore the searching process will be modified.

5.3.1 Searching with stemming

During the searching process the system attempts to retrieve as many relevant documents as possible. It attempts to do this in the following way.

- System receives a query with X terms
- For every X calculate its stem(S)
- For every term X retrieve the word id(W) of S
- For every W retrieve synonyms Y from the thesaurus
- For every term X retrieve the document ids from both the title- and abstract tables of the documents that contain W or Y.
- Take the Intersection of all the documents retrieved for every X.

At the end of this process the system will have a list of all documents containing all the stemmed terms and/or their stemmed synonyms.

5.4 Ranking

In this section the implementations of both the ranking types discussed in section 6.4 are discussed.

5.4.1 Ranking type 1

Ranking of type 1 adds a way to rank documents based on the frequency in which terms occur in documents. In order to keep track of these frequencies the indexing process will have to be adjusted. First the database design will be expanded so it can keep track of the frequency of terms. The new database schema is shown in figure 5.4.

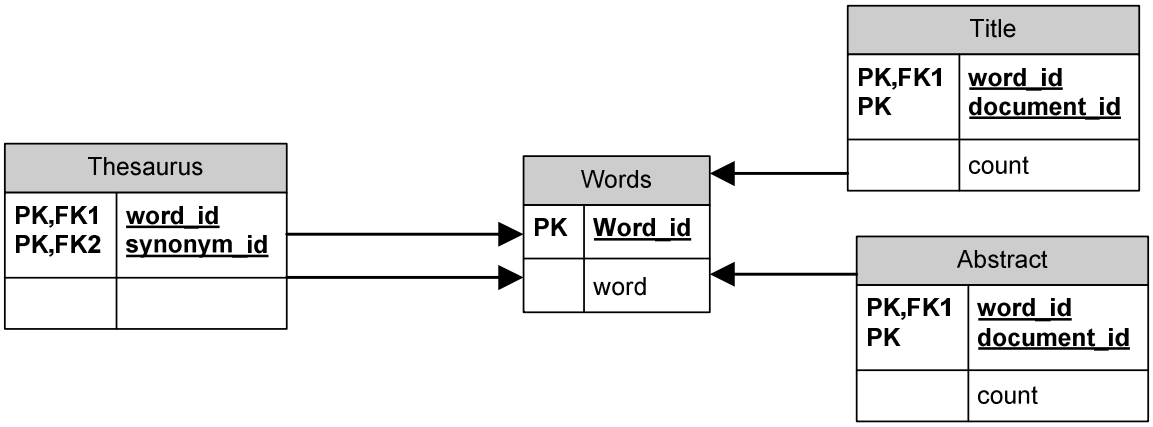


Figure 5.4 – Database schema including frequency of terms

Figure 5.4 shows that for every document and word the system now counts how many times a word occurs in either the title or the abstract.
The searching process remains unchanged but Instead of just returning a list of documents that match all the terms, the search process will now return a list of all documents that match the terms, but will also return the amount of times every term occurs in the title and/or abstract.

After the search returns its results the ranking process (section 5.1.4) starts. The ranking process will award 3 points to every hit of a term in the title and 2 points for every hit in the abstract. After the ranking process finishes ranking all documents, it will order the list of documents according to rank and present the results to the user.

5.4.2 Ranking type 2

Ranking type 2 not only keeps track of the frequency in which a word occurs in a title and/or abstract. It also keeps track how frequently a word occurs in separate documents. In order to do so a frequency counter has to be added to the words table. The new database schema is shown in figure 5.5.

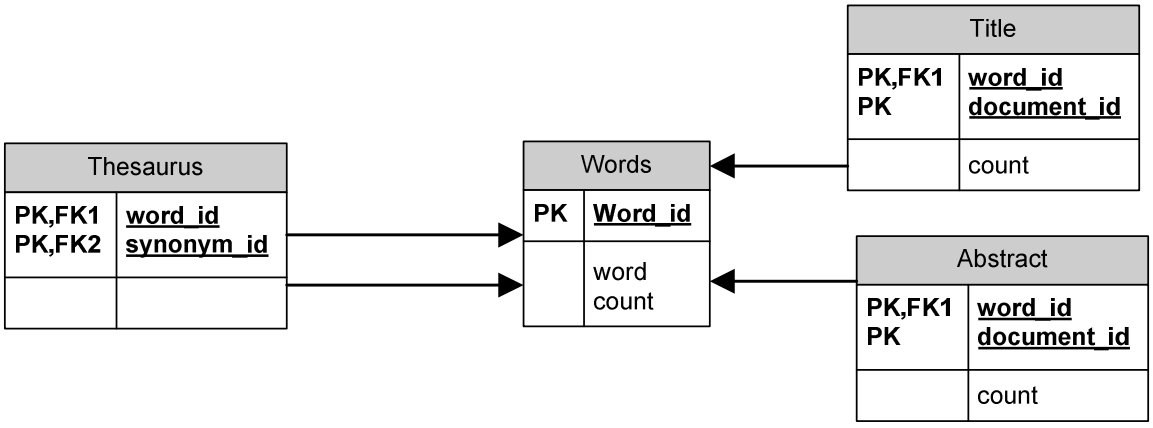


Figure 5.5- database schema including word frequencies.

The searching process remains the same was with searching of type 1.

The ranking system ranks the documents according to the formulas given in section 4.4.2. In order do so it occasionally has to query the words and/or thesaurus tables in order to check which words should be linked together as one term (because they are synonyms) and to check the frequency of a term in the words table. At the end the documents are ordered by rank again and presented to the user.

5.5 Multi-word terms

Instead of treating separate terms in multi-word terms as different terms, the system can attempt to guarantee that the words are in close proximity of each other before the terms are treated as a hit. The system will have to keep track of the location of a word in a title and/abstract. In order to do so the title and abstract tables will have to be expanded. The new database schema is shown in figure 5.6.

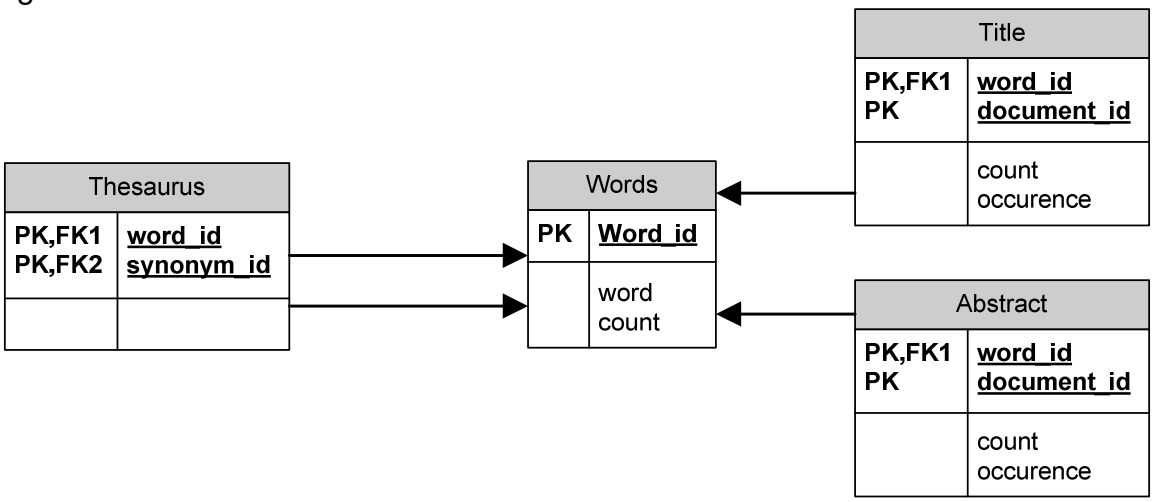


Figure 5.6 – Database schema including Multi-word support.

Figure 5.6 shows that both the title and abstract tables have a new column called "occurrence". This column contains an array of positions at which a word occurs in the text.

Searching the database has to be changed a little, it now has to retrieve the list of relevant documents including the count and occurrence of each search term.

During the ranking process the actual checking of multi-word terms is done. The ranking engine will analyze each sub-term of the multi-word term and check if it is in close enough proximity of the other sub-terms. If it is not close enough, the system will remove the term from its results (effectively removing it from the occurrence list at the index where it used to be and decrease the term counter by one). After all terms have been checked, the ranking engine will start its usual ranking routine, ranking all the separate terms found.
This means that the system can remove hits that are not in close enough proximity of each other, but the system can not rank documents based on multi word terms. The system ranks documents based on the individual sub-terms of the multi-word term. This is not a great loss, because the main reason for using multi-word terms is to remove the amount of irrelevant documents found. In theory it is possible to rank documents based on their multi-word terms this would however complicate the search process greatly, for this reason it was decided to rank documents based on the individual sub-terms of the multi-work terms.

5.6 mesh terms

All utilities in the previous sections only searched the title and the abstract of a document. In this section we will add the ability to search the mesh terms of the documents. In order to be able to do so a new table has to be added to the database design. The new database design is shown in figure 5.7.

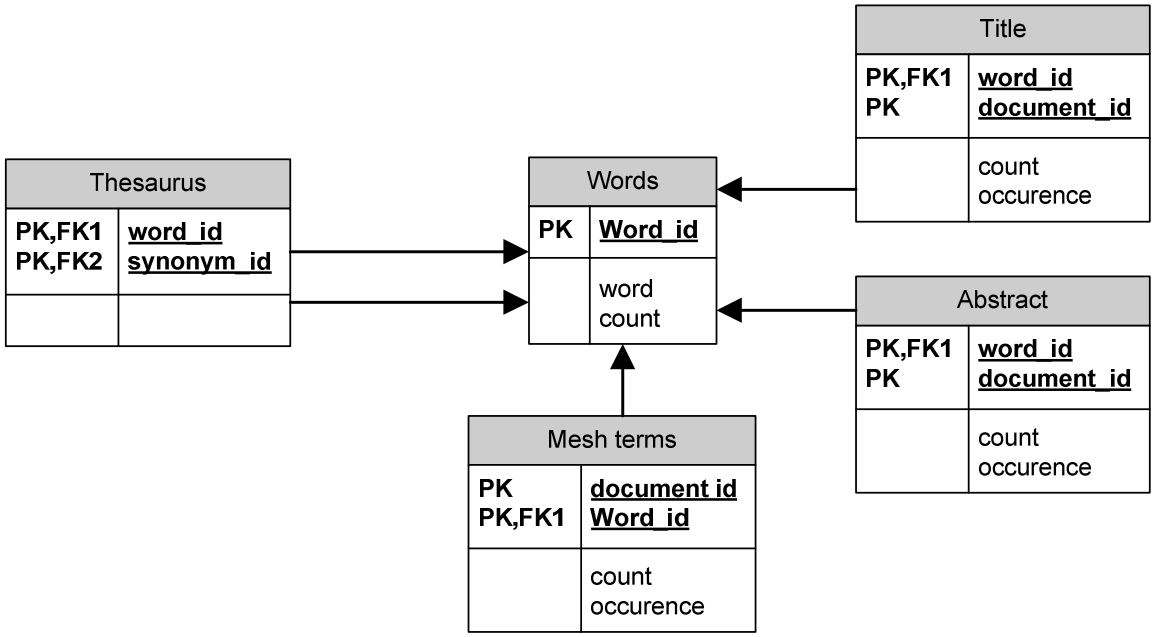


Figure 5.7 – Database schema including mesh terms

Figure 5.7 shows that the mesh term table is identical to the title and abstract tables. Indexing, searching and ranking are all the same on the mesh term table as on the title and abstract table. When using ranking type 1 the value of a "hit" in the mesh terms table is worth 3 points (like a hit in the title).

5.7 Cross-source searching

With the database design as shown in Figure 5.7 there is a limitation, it is impossible to search for different search terms across different tables. In order to make it possible for the system, to search for all the different information about a document, the information has to be stored in one table and not in three separate tables. The modified database schema is shown in figure 5.8

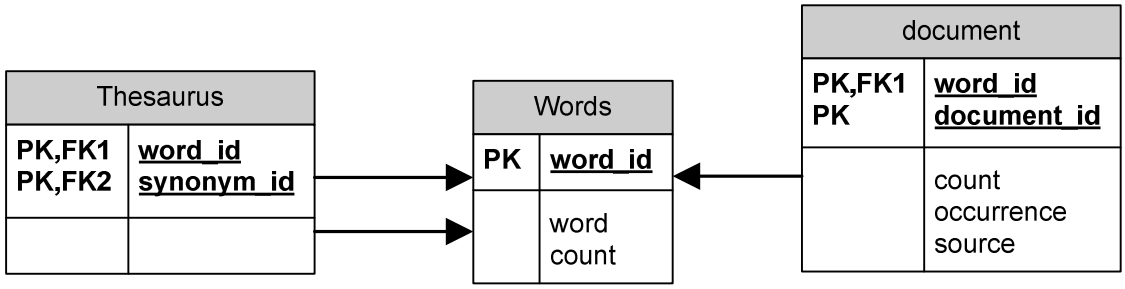


Figure 5.8 – Database schema with combined information

When comparing figures 5.7 and 5.8 it can be seen that the mesh term, abstract and title tables have been combined in a document table. The document table is identical to the tables it replaces but has one addition; it contains a source field. The source field contains what kind of information is stored in the record. Possible sources are mesh, title and abstract.

5.8 Query terms

In section 3.2 the fifty topics provided by TREC were divided into 3 different types. In this chapter we will decide on the terms that will be used when searching for every topic.

There is not a strict set of rules that dictates which terms to use. Deciding on the terms is something that is done by human judgment, but the terms are roughly decided upon like this:

- Type 1
  - Look for terms A and B
- Type 2
  - Look for C
- Type 3
  - Look for D and E

At this point there is no difference between how queries of type 1 and type 3 are handled. In section 7.4 a different set of query terms will be selected to improve performance, by then the difference between type 1 and type 3 will become apparent.

A complete list of terms that will be used can be found in appendix IX.

6 Judgment criteria

In order to test the effectiveness of our system, the results will be run though the TREC evaluation (*"trec_eval")* program [010]. This program generates information about the performance of a system.

The program takes two input files:

- The results of a run
- A relevance file.

The relevance file is provided by TREC to all its participants. It contains a list of relevant documents for all the topics TREC provided. A complete explanation of how this list was created can be found in [011].
Once the TREC evaluation program has judged a run, the following numbers will be looked at:

- Percentage found (Recall): this number represents the amount of relevant documents found (a Recall of 0,45 means that 45% of all relevant documents were found). This number is not affected by the amount of irrelevant documents retrieved, nor is it affected by the ranking documents have. This number only gives information about the recall of a run.
- Precision at 10 (P10): this number represents the number of relevant documents found in the 10 highest ranked documents. This number only gives information about the precision of a run.
- Precision at 100 (P100): this number represents the number of relevant documents found in the 100 most highly ranked documents. This number only gives information about the precision of a run.
- Mean average precision (MAP): The average precision of a single query is the mean of the precision scores after each relevant document retrieved. This number provides information about both the recall and the precision of a run.

More information about various measurements can be found in [005].

The TREC evaluation program provides several other numbers to measure the precision of a query. These will not be used. P10, P100 and MAP give a good enough view of the precision to compare the different methods. If in the future slight adjustments to the ranking system have to be tested, it is advised to also look at different measures.

The TREC program has one limitation: It can only accept 1.000 documents for every topic. If more documents are submitted then only the first 1.000 documents will be judged. If relevant documents are omitted because they are not in the first 1.000 documents then this will lower recall values (precision isn't affected because the documents that are omitted are submitted with a very low rank anyway).

7 Results

In order to research the actual effectiveness of the methods discussed in the previous chapters several tests were conducted. For every consecutive run one feature was added or removed to see the effectiveness of that feature. Table 7.1 shows which features were turned on at each individual run. The columns in the table represent a feature discussed in the previous chapters. The difference between “query term one” and “query term two” has not been discussed in previous chapters the difference between the two types will be discussed later in this chapter. The time listed, for each run, is the time needed to complete all 50 queries.

	Ranking	Stemming	Stop words	thesaurus	Query	Time	multi word terms	Mesh terms	Cross source searching
run1	no	No	no	No	term 1	230s	No	No	No
run2	no	No	no	Yes	term 1	460s	No	No	No
run3	no	No	yes	Yes	term 1	460s	No	No	No
run4	no	No	yes	Yes	term 2	285s	No	No	No
run5	no	Yes	yes	yes	term 2	400s	No	No	No
run6	type 1	Yes	yes	Yes	term 2	3100s	No	No	No
run7	type 1	Yes	yes	Yes	term 2	4200s	Yes	No	No
run8	type 2	Yes	yes	Yes	term 2	4300s	Yes	No	No
run9	type 1	Yes	yes	Yes	term 2	1800s	Yes	Yes	No
run10	type 1	Yes	yes	Yes	term 2	6100s	Yes	Yes	No
run11	type 1	Yes	yes	Yes	term 2	6100s	Yes	Yes	Yes

Table 7.1 the test features

A complete table of all the results can be found in appendix V up to appendix VIII. The tables show the four criteria (Recall, P10, P100 and MAP), as discussed in the previous chapter, for the individual topics and for all topics as a whole.

As can be seen in table 7.1 not all possible combinations of features were tested. This is because it was decided to apply new features to the best run up to that moment. E.g. to test the effectiveness of stemming the best run (not using stemming) was selected and stemming was added. This way all features can be tested in a minimal amount of runs. The result of this method of testing is that there is usually just one difference between the rows in the table.

7.1 The database

All the runs that were conducted were run on the same database server, and the configuration of that server did not change between the different runs. All the tables that were used had indexes on all their columns. The abstract table had about 380 million records and the title table had about 50 million records. Both tables were clustered on the hard disk on their word id column, to improve performance. Besides the clustering function no specific postgreSQL functions were used.

The total storage space used by the database when including stop words is 22050 MB (the table) + 10393 MB (indices) = 32443 MB. The total storage space used by the database when removing stop words is 19380 MB (the table) + 7125 MB (indices) = 26505 MB.

7.2 Thesaurus

Table 7.2 shows an overview of all the runs, and the features used on those runs, that will be discussed in this section.

	Ranking	Stemming	Stop words	thesaurus	Query	Time	multi word terms	Mesh terms
run1	no	No	no	No	term 1	230s	No	No
run2	no	No	no	Yes	term 1	460s	No	No

Table 7.2 Overview of the relevant runs.

The effectiveness of the thesaurus was tested between runs one and two. The full results of those runs can be found in appendix V up to appendix VIII but an abbreviation of the results is given in table 7.3.

Recall		**Map**		**P10**		**P100**	
run1	*run2*	*run1*	*run2*	*run1*	*Run2*	*run1*	*Run2*
0,27	0,28	0,12	0,10	0,33	0,26	0,18	0,17

Table 7.3 the overall effect of the thesaurus

As can be seen in table 7.3 the recall goes up slightly but the precision goes down on all 3 measurements for it. This was to be expected, but to get a better look at the effectiveness of the thesaurus an analysis of all the topics was conducted. The results of the analysis between the first and second run are shown in table 7.4.

Thesaurus effectiveness	Recall	P10	P100	Map
Improved	22%	4%	14%	12%
Unchanged	78%	78%	78%	74%
Decreased	0%	18%	8%	14%

Table 7.4 the percentage of topics that change when using the thesaurus

Table 7.4 can tell us something about the usefulness of the thesaurus. When looking at recall it is clear that the thesaurus just improves it and never lowers it. Precision on the other hand is another story. In table 7.3 we see that all 3 measurements decline yet when looking at table 7.4 we see that P100 actually improves on more topics then declines. Based on these numbers no claims can be made about the effect of the thesaurus on the precision, but for recall the thesaurus is a definite improvement.

Based on these numbers it was decided to use the thesaurus in all further runs (with the exception of run 5 the reason for this will be discussed later).

7.3 Stop words

Table 7.5 shows an overview of all the runs, and the features used on those runs, that will be discussed in this section.

	Ranking	Stemming	Stop words	thesaurus	Query	Time	multi word terms	Mesh terms
run2	no	No	no	Yes	term 1	460s	No	No
run3	no	No	yes	Yes	term 1	460s	No	No

Table 7.5 Overview of the relevant runs.

To get an overview of the effect the removal of stop words has on the results a closer look has to be taken at runs two and three. The full results of those runs can be found in appendix V up to appendix VIII but an abbreviation of the results is given in table 7.6.

Recall		**Map**		**P10**		**P100**	
run2	*run3*	*Run2*	*run3*	*run2*	*run3*	*run2*	*Run3*
0,28	0,28	0,10	0,10	0,26	0,26	0,17	0,17

Table 7.6 the overall effect of the removal of stop words

As can be seen in table 12.4 there is no change in any of the measurement criteria. A closer took at all the separate topics also shows that there is no change at any of the topics.

The removal of stop words was primarily added to the system to reduce the time needed to answer a query, while not changing the recall and precision. Based on the results we can conclude that the results are not influenced by this feature.

Looking at the speed of the system we can see that removing stop words has no significant influence on the execution time, run two and three both have an execution time of 460 seconds (table 7.5).

This feature has no significant influence on the performance of the system, yet it was added to all the runs following run three. This was done because this feature does reduce the amount of storage space needed by the system (as discussed in section 7.1).

7.4 Evaluation of the query terms

Table 7.7 shows an overview of all the runs, and the features used on those runs, that will be discussed in this section.

	Ranking	Stemming	Stop words	thesaurus	Query	Time	multi word terms	Mesh terms
run3	no	No	yes	Yes	term 1	460s	No	No
run4	no	No	yes	Yes	term 2	285s	No	No

Table 7.7 Overview of the relevant runs.

After looking at the recall values of runs one up to three there was some disappointment about the recall values the system achieved. In order to improve recall it was decided to lower the amount of terms the system uses to try and find documents.

This change probably lowers the precision, but in order to ever become more precise it has to have a good recall value. The precision will be dealt with using features discussed in sections following this one.

Of the three topic types discussed in chapter eight roughly the following changes were made:

- Type 1: Remove term B from the query terms
  - This is because term B isn't always in the title/abstract but can be in the document itself. Since term A has far great significance B can be dismissed.
- Type 2: Remove references to a protocol, method or function from the terms
  - Because a reference to a protocol, method or function might be omitted in a title and/or abstract.
- Type 3: No changes
  - Because both terms are equally important.

As with the initial list of terms this list was made using human judgment, this means not all changes follow the three rules listed above. For every topic the amount of retrieved documents together with recall were evaluated. If the results were below average then the rules mentioned above would be applied. The updated list with query terms can be found in appendix X.

The effectiveness of the new query terms can be seen when comparing run three and four. Table 7.8 gives an overview of the results of runs three and four.

Recall		**Map**		**P10**		**P100**	
run3	*Run4*	*run3*	*run4*	*run3*	*run4*	*run3*	*run4*
0,28	0,36	0,10	0,14	0,26	0,23	0,17	0,17

Table 7.8 the overall effect of changing the query terms

When looking at table 7.8 it becomes clear that recall went up but also MAP has improved. At first glance this looks promising but a more thorough analysis of all 50 topics involved is required. The results of this analysis can be found in table 7.9.

Change of terms effect	Recall	P10	P100	Map
Improved	30%	16%	30%	26%
Unchanged	58%	66%	52%	54%
Decreased	12%	18%	18%	20%

Table 7.9 the percentage of topics that change when using different query terms

The first thing that comes to mind when looking at table 7.9 is the fact that recall goes down at 12% of the topics. This looks odd since loosening query terms should not make the system “loose” relevant documents. In this case this can happen because of the limitation which TREC put on the amount of documents participants can submit on every topic (a maximum of 1.000 documents for every topic). When removing to many search terms the maximum amount of documents is exceeded and documents are omitted from the results, if there are relevant documents in the omitted part then recall goes down.
Another fact that looks odd is the fact that precision actually goes up more then it goes down. But overall it is not affected as greatly as recall by the changes.

Changing the query terms has improved recall by almost 27% and has not influenced precision as badly as expected. Therefore the altered set of query terms will be used in all the runs following the sixth.

7.5 Stemming

Table 7.10 shows an overview of all the runs, and the features used on those runs, that will be discussed in this section.

	Ranking	Stemming	Stop words	thesaurus	Query	Time	multi word terms	Mesh terms
run4	no	No	yes	Yes	term 2	285s	No	No
run5	no	Yes	yes	yes	term 2	400s	No	No

Table 7.10 Overview of the relevant runs.

To see the effect of stemming we have to compare run four and five. The overall results from those runs are listed in table 7.11.

Recall		**MAP**		**P10**		**P100**	
run4	*Run5*	*run4*	*run5*	*run4*	*run5*	*run4*	*run5*
0,36	0.42	0,14	0.13	0,23	0.23	0,17	0.15

Table 7.11 the overall results of stemming

When looking at table 7.11 a 17% increase in recall and almost no change in precision is noticed. Again a more precise breakdown of the topics is made and can be found in table 7.12.

Stemming effect	Recall	P10	P100	Map
Improved	44%	18%	14%	34%
Unchanged	28%	66%	46%	26%
Decreased	28%	16%	40%	40%

Table 7.12 the percentage of topics that change when using stemming

Based on the numbers of table 7.12 it can be concluded that stemming increases recall more often then lowering it. Stemming does not seem to affect P10 and MAP in a definite way (increase is about as big as the decrease) yet it decreases P100 a lot more then it increases it.

Because of the increase in recall the stemming process was added to all runs following the eighth run.

7.6 Ranking

Table 7.13 shows an overview of all the runs, and the features used on those runs, that will be discussed in this section.

	Ranking	Stemming	Stop words	thesaurus	Query	Time	multi word terms	Mesh terms
run5	no	Yes	yes	yes	term 2	400s	No	No
run6	type 1	Yes	yes	Yes	term 2	3100s	No	No

Table 7.13 Overview of the relevant runs.

When it comes to ranking, two kinds of ranking were tested. First the effects of ranking will be compared to using no ranking at all (by looking at the difference between no ranking and ranking of type 1, as discussed in section 6.4.1). In the next section the difference between the two types of ranking will be discussed.

The effect of ranking can be seen by comparing the results of runs five and six. The overall results of those runs are listed in table 7.14.

Recall		**MAP**		**P10**		**P100**	
run5	*Run6*	*run5*	*run6*	*run5*	*run6*	*Run5*	*run6*
0,42	0,43	0,13	0,19	0,23	0,37	0,16	0,22

Table 7.14 the overall results of ranking

As can be seen in table 7.14 the system improves on all 4 measuring criteria. The improvement in precision (MAP, P10, P100) was expected but the improvement in recall was not expected. The improvement in recall can be explained though. When submitting results to TREC only the first 1000 results for a topic get processed. For several topics the system found more then 1000 results. By submitting only the highest ranked documents the system managed to get a higher recall by using ranking. Table 7.15 shows the results of ranking on all topics.

Ranking effect	Recall	P10	P100	Map
Improved	20%	34%	54%	68%
Unchanged	80%	60%	42%	26%
Decreased	0%	6%	4%	6%

Table 7.15 the percentage of topics that change when using ranking

As can be seen in table 7.15 recall improves in 20% of the topics but never decreases. Looking at precision we see that the system improves on a lot more topics then that it decreases.

Bases on the number from tables 7.14 and 7.15 it is concluded that ranking can be used to increase both recall and precision.

7.7 Ranking types

Table 7.16 shows an overview of all the runs, and the features used on those runs, that will be discussed in this section.

	Ranking	Stemming	Stop words	thesaurus	Query	Time	multi word terms	Mesh terms
run7	type 1	Yes	yes	Yes	term 2	4200s	Yes	No
run8	type 2	Yes	yes	Yes	term 2	4300s	Yes	No

Table 7.16 Overview of the relevant runs.

The results of the two types of ranking, as discussed in section 6.4, that were used can be evaluated by looking at runs seven and eight. Table 7.17 shows the overall results of both ranking types.

Recall		**MAP**		**P10**		**P100**	
run7	*run8*	*run7*	*run8*	*run7*	*run8*	*run7*	*run8*
0,44	0,44	0,20	0,20	0,42	0,42	0,23	0,24

Table 7.17 the overall results of the two ranking types

As expected the recall value between the runs does not change. When looking at precision the effect of the ranking on the topics is insignificant. In order to get a better look at the effects an evaluation of all topics is needed. The results of this evaluation are shown in table 7.18.

Ranking type effect	Recall	P10	P100	Map
Improved	0%	20%	20%	40%
Unchanged	100%	58%	66%	18%
Decreased	0%	22%	14%	42%

Table 7.18 the percentage of topics that change when using ranking type 2

The evaluation of the topics shows that both types of ranking perform at about the same level. It is impossible to tell which type performs better.

7.8 Multi-word-terms

Table 7.19 shows an overview of all the runs, and the features used on those runs, that will be discussed in this section.

	Ranking	Stemming	Stop words	thesaurus	Query	Time	multi word terms	Mesh terms
run6	type 1	Yes	yes	Yes	term 2	3100s	No	No
run7	type 1	Yes	yes	Yes	term 2	4200s	Yes	No

Table 7.19 Overview of the relevant runs.

When using multi-word-terms it was decided to only use these terms in the ranking engine of the system. This means that the system still looks for the individual words, of the terms it gets, but when these words are not close to each other the ranking engine wont reward “points” to it. This way the recall values of the system should remain the same or better but the precision should go up.

The effect of using multi-word-term support can be evaluated by looking at run six and seven. The overall results of those runs are shown in table 7.20.

Recall		**MAP**		**P10**		**P100**	
Run6	*run7*	*run6*	*run7*	*Run6*	*run7*	*run6*	*run7*
0,43	0,44	0,19	0,20	0,37	0,42	0,22	0,23

Table 7.20 the overall results of multi-word-terms

At first glance the system looks to have improved on all accounts. A closer evaluation of all the topics is shown in table 7.21.

Ranking type effect	Recall	P10	P100	Map
Improved	6%	22%	24%	34%
Unchanged	94%	72%	68%	52%
Decreased	0%	6%	8%	14%

Table 7.21 the percentage of topics that change when using multi-word terms

Table 7.21 shows that at far more topics the precision increased then decreases. Based on these numbers it can be concluded that multi-word terms can be used to increase precision of the system.

7.9 comparison of query types

In section 5.2 all fifty topics were divided into three types. In this section we will analyze the performance of the system on all of these types. To compared these types the average of the recall, MAP, P10 and P100 will be used to see how they perform. To compare the different types we will look at run ten because that run had the best overall performance. Table 7.22 shows the average results of the three query types.

Overall effect	Avg. Recall	Avg. P10	Avg. P100	Avg. MAP
Type 1	0,41	0,41	0,19	0,20
Type 2	0,42	0,45	0,28	0,21
Type 3	0,63	0,13	0,13	0,11

Table 7.22 the average results of the query types.

Table 7.22 shows us that type one and two have about the same recall. Type two has a slight advantage over type one, on both recall and MAP, but the difference is to small too make an impression. On P10 and P100 however type two clearly beats type one on precision. Type three on the other had has a big advantage over types one and two on recall and a big disadvantage on precision, but since there are only four topics of type three it is to soon conclude how the system performs on topics of type three more topics of that type would have to be available to make a proper evaluation of the results. Based on the numbers in table 7.22 the system performs best on queries of type two.

7.10 Mesh terms

Table 7.23 shows an overview of all the runs, and the features used on those runs, that will be discussed in this section.

	Ranking	Stemming	Stop words	thesaurus	Query	Time	multi word terms	Mesh terms
run7	type 1	Yes	yes	Yes	term 2	4200s	Yes	No
run9	type 1	Yes	yes	Yes	term 2	1800s	Yes	Yes
run10	type 1	Yes	yes	Yes	term 2	6100s	Yes	Yes

Table 7.23 Overview of the relevant runs.

In order to test the effectiveness of including mesh terms into the search space, two test runs were conducted. Run9 was made on just the mesh terms data. This means the system did not look at the title and abstract of the documents. A summary of the results of run nine is shown in table 7.24.

Recall	**MAP**	**P10**	**P100**
Run9	*Run9*	*Run9*	*Run9*
0.16	0.04	0.16	0.11

Table 7.24 the results of searching the mesh terms alone

Table 7.24 shows that the performance of searching just the mesh terms doesn't give a high precision but at recall of 16% is still quite high considering we are just looking at the mesh terms.

After the results of run9 it was decided to do another run combining the mesh terms, the title and the abstracts of documents. To get a good view on the effect of this addition we have compare run seven and ten. Table 7.25 shows the comparison between run seven and run ten.

Recall		**MAP**		**P10**		**P100**	
Run7	*run10*	*Run7*	*run10*	*Run7*	*run10*	*Run7*	*run10*
0,44	0.43	0,20	0,20	0,42	0,40	0,23	0,24

Table 7.25 the results of searching in mesh, title and abstract

Judging from table 7.25 including mesh terms doesn't seem to have any effect on the system. A close comparison of both runs is shown in table 7.26.

Mesh term inclusion effect	Recall	P10	P100	Map
Improved	16%	16%	4%	18%
Unchanged	64%	60%	90%	76%
Decreased	20%	24%	6%	6%

Table 7.26 the percentage of topics that change when using mesh terms

Table 7.26 also shows that the effects of including mesh terms are minimal.

7.11 Cross-source searching

Table 7.27 shows an overview of all the runs, and the features used on those runs, that will be discussed in this section.

	Ranking	Stemming	Stop words	thesaurus	Query	Time	multi word terms	Mesh terms	Cross source searching
run10	type 1	Yes	yes	Yes	term 2	6100s	Yes	Yes	No
run11	type 1	Yes	yes	Yes	term 2	6100s	Yes	Yes	Yes

Table 7.27 Overview of the relevant runs.

In this section the effects of combining all content into one table will be discussed. In order see the effect of using cross-source searching runs 10 and 11 have to be compared. Table 7.28 shows the results of runs 10 and 11.

Recall		**MAP**		**P10**		**P100**	
Run10	*run11*	*Run10*	*run11*	*Run10*	*run11*	*Run10*	*run11*
0,43	0.47	0,20	0,20	0,40	0,35	0,24	0,23

Table 7.28 the results of searching in mesh, title and abstract

Table 7.28 shows a slight increase in recall but also a slight decrease in precision. Table 7.29 shows the change in recall and precision over all topics.

Cross-source searching effects	Recall	P10	P100	Map
Improved	50%	36%	32%	38%
Unchanged	34%	42%	32%	12%
Decreased	16%	22%	36%	50%

Table 7.29 the percentage of topics that change when using mesh terms

Table 7.29 shows that recall is better or unchanged on 84% of the topics, based on this it can be concluded that using cross-source searching does indeed improve recall slightly. Precision however increases and decreases in about the same rate when looking at different topics, based on this no real comments can be made about the influence of cross-source searching on the overall precision of the system.

8 Evaluation of the results

The results as described in the previous chapter were mostly as expected but some of the results were unexpected. In this chapter we will try to explain these unexpected results.

8.1 Why is the effect of the thesaurus so limited?

Before the testing runs began it was expected that the thesaurus would improve recall most of the features that would be added. Yet the results show only a small improvement in recall when using a thesaurus. After evaluation of the topics and the results the reason why the increase was so small was found.

As discussed in section 4.1 only ten of the topics utilize the thesaurus. When only the ten topics that use the thesaurus are considered the effects of it are more like expected. Table 8.1 shows the average Recall, P10, P100 and MAP for the ten topics that use the thesaurus.

thesaurus	recall	p10	p100	map
Without	0,37	0,40	0,19	0,18
With	0,50	0,22	0,22	0,17

Table 8.1 the average number of the topics using the thesaurus.

If we look at the effect of the thesaurus on the recall then we see a big increase, like expected before the testing runs started. Also as expected the precision of the system goes down when using a thesaurus.

8.2 Why is there little difference between the ranking types?

Before the testing runs were completed it was expected that a ranking system of type two (the altered TF*IDF ranking) would outperform the basic ranking that was in place just for testing purposes. Yet the results show that both ranking mechanisms perform at the same efficiency. After evaluation of all the topics the reason was found for this result.

Ranking of type one performs the same as type two because the query terms were already decided upon by hand. During this process the person selecting the query terms already removed all irrelevant data, and thus most of the documents that would receive a poor ranking.

Ranking type two values the search terms based on the amount of times a term occurs in the dataset. This is a good way for the system to differentiate between important and unimportant query terms. This however works best when the

system is presented with a lot of query terms which are not all of the same importance. In the current system the query consist of just one, two or three query terms and all of them are equally important. Meaning they will all be awarded a “high” rank by the TF*IDF ranking. This is exactly what the basic ranking system does.

In order for the ranking type two to really shine the system would have to be adjusted so it would look for all the terms in the topics (the full topic description not just the keywords that were selected) and not use the intersection but the union of the documents found for the separate terms. This way the ranking system could really differentiate between the query terms. This would also remove the only human step needed to search the database for the documents required by the topics since the database would rank the terms of the topic automatically by importance. This would however greatly increase the amount of resources (and thus time) needed by the system to find results for all topics.

8.3 Why is the effect of searching mesh terms so limited?

It was expected that including the mesh terms for searching would improve both precision and recall of the system. If there was any effect when searching, then it was a decrease in precision. After some consideration about how mesh terms are generated a possible reason for this was found.
On average documents have ten mesh terms defined on them. These mesh terms are very vague about the contents of an article, because there is not enough room to create a complete overview of the document (this is what the abstract is for).
In order to use these mesh terms effectively they have to be combined with the title and the abstract of the document. Instead of searching through the three fields in the dataset separately they should be treated as on piece of information. This way the system will have a lot more information to process.
If the system is searching for three terms it can not find documents with one term in the title, one in the abstract and one in the mesh terms at this moment. This greatly limits the usefulness of the mesh terms.

8.4 Why is the overall recall only 47%?

Before the testing began the expectation of the total recall that was achievable was higher. A closer look at the dataset revealed the reason why recall could not get as high as expected. Of all the 3.5 Million documents in the database only about 65% (determined by taking randomly 10.000 documents and looking if an abstract was defined) of them has an abstract defined for them. This means that with the fields that are currently used to search the database 35% of the documents was considered just by looking at the title and the mesh terms. This together with the issue discussed in the previous section causes the recall to be lower then expected.

8.5 Why is the effect of cross-source searching so limited?

At first the support for cross-source searching was expected to give a big increase in both recall and precision. Yet the results in section 7.11 show that there was just a slight increase in recall and close to no change in precision. This has two reasons.

- The use of a limited amount of search terms.
  Because the topics all have a very limited amount of search terms the change that just one or two of them are missing from one of the sources and actually occurring in another source (while still relevant) are very limited. As with the ranking systems discussed in section 8.2 using cross-source searching would probably be more useful when searching for more search terms.
- Relevance of sources that contain all the search terms is far bigger then sources that contain just a subset of the search terms.
  A “hit” in the title that contains all the searched terms is far more likely to be relevant then a hit on the title with just a subset of the search terms, and the rest of the terms in either the mesh terms or the abstract.

9 Comparison with others

In this chapter a comparison will be made between the results as described in the previous chapter and the results of other TREC participants. The results will be compared to the 47 runs that were submitted to TREC (these 47 runs do not include the runs made for this report).

9.1 General comparison

In this section our results will be compared to all the runs submitted to TREC. Table 9.1 shows the results of both our system and the mean of the 47 runs that were submitted. The overall results of all the 47 submitted runs to TREC can be found in [011].

	Overall MAP	**Overall P10**	**Overall P100**
Our system	0,20	0,42	0,24
Mean	0,21	0,45	0,26

Table 9.1 Comparison between our results and the mean of the runs submitted.

As can be seen in table 9.1 our system performs about average compared to the 47 runs submitted to TREC. In order to get a better look at the results a comparison was made between the results of every single topic. Because the mean Map, P10 and P100 were not available for the separate topics we used the median MAP, P10 and P100 to determine how we performed on the separate topics. The results for very topic can be found in appendix XI.

	MAP	P10	P100
Better then median	34 %	30 %	30 %
The same as median	8 %	22 %	6 %
Worse then median	58 %	48 %	64 %

Table 9.2 the performance of our system compared to the median of the 47 submitted runs

As can be seen in table 9.2 our system performs slightly below median on average.

9.2 Comparison against the best run

The best run was submitted by Patolis Corp. and their results are listed in table 9.3

	Overall MAP	**Overall P10**	**Overall P100**
Best	0,41	0,60	0,42

Table 9.3 the results of Patolis Corp.

Table 9.3 shows that Patolis corp. achieved about double the MAP and precision our system achieved. In this section we will try to compare both systems and try to explain the difference in performance.

First of all there all there are a few things both systems have in common. Both are based on the PostgreSQL database system, and both use the porter stemmer. Both also use some form of TF*IDF ranking and remove stopwords.

Besides the things both systems have in common Patolis corp. made a few additions. They use the following additions:

- Pseudo-Relevance feedback: The system adjusts the terms it is looking for based on the terms, which are found in the documents with the highest rank by looking just at the terms provided.
- Reference database feedback from both the LocusLink summary and the MeSH entry database: this works as a thesaurus. For every term that is being searched for the system tries to find synonyms and abbreviations from both locuslink and the Mesh entry database.
- Smoothing: the process adjusts the maximum likelihood estimator of a language model, so that it will be more accurate. The smoothing process is quite complicated, a full overview over smoothing can be found in [016].
- Document-dependant priors: based on the amount of information that is available on a document predictions can be made about the likelihood of a document being relevant. Taking documents into consideration; the less information available, the more likely the document is relevant.

A more thorough description of the Patolis corp. runs can be found in [012].

10 Conclusions and future improvements

Of all the features that were tried only the removal of stop words had close to no impact on the system. Removing stop words has only an impact on the amount of storage space.

The use of a thesaurus can greatly improve recall, if topics actually use them, but has a slightly negative effect on precision. Overall the increase in recall outweighs the decrease in precision.

Stemming has a comparable effect as a thesaurus meaning it increases recall but slightly reduces precision. Again the increase in recall outweighs the decrease in precision.

Ranking has close to no effect on recall but greatly increases precision. The difference between ranking types could not be tested sufficiently to really tell what kind of ranking is best.

The use of multi-word-terms slightly improves precision at no cost of recall.

Given the current setup of the system, the system just looking at the title, abstract and mesh terms on a desktop system running an out of the box database system, the result is very promising. The system works reasonably fast and the recall and Precision at 10 and 100 are very workable.

The ranking types that were tested were not tested to their fullest extent. In order to test the effects of the TF*IDF ranking the method of querying has to be changed so that the system can take full sentences directly from the topics. This can be done by making all search terms optional instead of mandatory in the results. Changing this could remove the last step of human judgment in the searching process and could also increase precision quite a lot.

11 References

[001] Baeza-Yates, Ribeiro-Neto.
Modern Information Retrieval.
ACM press 1999

[002] William R. Hersh.
Information Retrieval: a Health and Biomedical Perspective – second edition.
Springer 2003 – pages 170-190

[003] G G Chowdhury
Introduction to modern information retrieval
Library association publishing, 1999

[004] David A. Grossman, Ophir Frieder
Information retrieval Algorithms and heuristics
1998

[005] Chris Buckley, Ellen M. Voorhees
Retrieval Evaluation with Incomplete Information
SIGIR 2004

[006] http://www.ncbi.nih.gov/entrez/query.fcgi?db=gene

[007] Porter M.
An algorithm for suffix stripping
1980

[008] http://trec.nist.gov

[009] http://www.postgresql.org

[010] C. Buckley. trec eval IR Evaluation Package.
ftp://ftp.cs.cornell.edu/pub/smart.

[011] William R. Hersh, Ravi Teja Bhuptiraju, Laura Ross, Phoebe Johnson, Aaron M. Cohen, Dale F. Kraemer
TREC 2004 Genomics Track Overview
2004

[012] Sumio FUJITA
Revisiting Again Document Length Hypotheses TREC-2004 Genomics Track Experiments at Patolis
2004

[013] Salton G.
Automatic Text Processing
Addison-Wesley 1989

[014] Julie Beth Lovins
Development of a Stemming Algorithm
Mechanical translation and computational linguistics, 11:22-31
1968

[015] www.dictionary.com

[016] chengXiang Zhai and John lafferty
A Study of Smoothing Methods for Language Models Applied to Ad Hoc Information Retrieval
2001
[017] Kazuhiro Seki, James C. Costello, Vasanth R. Singan, and Javed Mostafa
TREC 2004 Genomics Track Experiments at IUB
Indiana University Bloomington, 2004
[018] Colleen Crangle, Alex Zbyslaw, J. Michael Cherry, Eurie L. Hong
Concept Extraction and Synonymy Management for Biomedical Information Retrieval
Stanford University, 2004
[019] Patrick Ruch, Frédéric Ehrler, Johan Marty, Christine Chichester, Gilles Cohen, Paul Fabry, Henning Müller, Antoine Geissbühler
Report on the TREC 2004 Experiment: Genomic Track
University Hospital of Geneva, 2004
[020] Alan R. Aronson,Dina Demner,Susanne M. Humphrey,Nicholas C. Ide,Won Kim,Hongfang Liu,Russell R. Loane,
James G. Mork,Lawrence H. Smith,Lorraine K. Tanabe,W. John Wilbur,Natalie Xiea
Knowledge-intensive and statistical approaches to the retrieval and annotation of genomics MEDLINE citations
National Library of Medicine, 2004

12 Appendices

Appendix I – Pubmed fields

In the table below you can see all possible information that could be used for one publication. All this and more information can be found at: http://www.ncbi.nlm.nih.gov/entrez/query/static/help/pmhelp.html#MEDLINEDisplayFormat

Tag	Name	Description
AB	Abstract	Abstract
AD	Affiliation	Institutional affiliation and address of the first author, and grant numbers
AID	Article Identifier	Article ID values may include the pii (controlled publisher identifier) or doi (Digital Object Identifier)
AU	Author	Authors
CI	Copyright Information	Copyright statement
CIN	Comment In	Reference containing a comment about the article
CN	Corporate Author	Corporate author or group names with authorship responsibility
CON	Comment On	Reference upon which the article comments
DA	Date Created	Used for internal processing at NLM
DCOM	Date Completed	Used for internal processing at NLM
DEP	Date of Electronic Publication	Electronic publication date
DP	Publication Date	The date the article was published
EDAT	Entrez Date	The date the citation was added to PubMed
EFR	Erratum For	Cites the original article needing the correction
EIN	Erratum In	Reference containing a published erratum to the article
FAU	Full Author Name	Full Author Names
FIR	Full Investigator	Full investigator name
FPS	Full Personal Name as Subject	Full Personal Name of the subject of the article
GN	General Note	Supplemental or descriptive information related to the document
GR	Grant Number	Research grant numbers, contract numbers, or both that designate financial support by any agency of the US PHS (Public Health Service)
GS	Gene Symbol	Abbreviated gene names (used 1991 through 1996)
IP	Issue	The number of the issue, part, or supplement of the journal in which the article was published
IR	Investigator	NASA-funded principal investigator

IRAD	Investigator Affiliation	Affiliation of NASA-funded principal investigator
IS	ISSN	International Standard Serial Number of the journal
JID	NLM Unique ID	Unique journal ID in NLM's catalog of books, journals, and audiovisuals
LA	Language	The language in which the article was published
LR	Last Revision Date	The date a change was made to the record during a maintenance procedure
MH	MeSH Terms	NLM's controlled vocabulary
MHDA	MeSH Date	The date MeSH terms were added to the citation. The MeSH date is the same as the Entrez date until MeSH are added
OAB	Other Abstract	Abstract supplied by an NLM collaborating organization
OCI	Other Copyright Information	Copyright owner
OID	Other ID	Identification numbers provided by organizations supplying citation data
ORI	Original Report In	Displays on Patient Summary. Cites original article associated with the patient summary
OT	Other Term	Non-MeSH subject terms (keywords) assigned by an organization identified by the Other Term Owner
OTO	Other Term Owner	Organization that provided the Other Term data
OWN	Owner	Organization acronym that supplied citation data
PG	Pagination	The full pagination of the article
PHST	Publication History Status Date	History status date
PL	Place of Publication	Journal's country of publication
PMID	PubMed Unique Identifier	Unique number assigned to each PubMed citation
PS	Personal Name as Subject	Individual is the subject of the article
PST	Publication Status	Publication status
PT	Publication Type	The type of material the article represents
RF	Number of References	Number of bibliographic references for Review articles
RIN	Retraction In	Retraction of the article
RN	EC/RN Number	Number assigned by the Enzyme Commission to designate a particular enzyme or by the Chemical Abstracts Service for Registry Numbers
ROF	Retraction Of	Article being retracted
RPF	Republished From	Original article
RPI	Republished In	Corrected and republished article
SB	Subset	Journal/Citation Subset values representing various topic areas
SFM	Space Flight Mission	NASA-supplied data space flight/mission name and/or number
SI	Secondary Source Identifier	Identifies a secondary source that supplies information, e.g., other data sources, databanks and accession numbers of molecular

		sequences discussed in articles
SO	Source	Composite field containing bibliographic information
SPIN	Summary For Patients In	Cites a patient summary article
STAT	Status Tag	Used for internal processing at NLM
TA	Journal Title Abbreviation	Standard journal title abbreviation
TI	Title	The title of the article
TT	Transliterated / Vernacular Title	Non-Roman alphabet language titles are transliterated.
UIN	Update In	Update to the article
UOF	Update Of	The article being updated
VI	Volume	Journal volume

Appendix II – The Topics

<ID>1</ID>
<TITLE>Ferroportin-1 in humans</TITLE>
<NEED>Find articles about Ferroportin-1, an iron transporter, in humans.</NEED>
<CONTEXT>Ferroportin1 (also known as SLC40A1; Ferroportin 1; FPN1; HFE4; IREG1; Iron regulated gene 1; Iron-regulated transporter 1; MTP1; SLC11A3; and Solute carrier family 11 (proton-coupled divalent metal ion transporters), member 3) may play a role in iron transport.</CONTEXT>

<ID>2</ID>
<TITLE>Generating transgenic mice</TITLE>
<NEED>Find protocols for generating transgenic mice.</NEED>
<CONTEXT>Determine protocols to generate transgenic mice having a single copy of the gene of interest at a specific location.</CONTEXT>

<ID>3</ID>
<TITLE>Time course for gene expression in mouse kidney</TITLE>
<NEED>What is the time course of gene expression in the murine developing kidney?</NEED>
<CONTEXT>Relevant articles describe genes involved in kidney development.</CONTEXT>

<ID>4</ID>
<TITLE>Gene expression profiles for kidney in mice</TITLE>
<NEED>What mouse genes are specific to the kidney?</NEED>
<CONTEXT>What genes are expressed only in the mouse kidney and not in other tissues?</CONTEXT>
</TOPIC>

<ID>5</ID>
<TITLE>Protocols for isolating cell nuclei</TITLE>
<NEED>Articles are relevant if they describe methods for subcellular fractionation of nuclei.</NEED>
<CONTEXT>Laboratory preparations can be enriched for certain kinds of proteins if the cellular compartment in which they reside is purified away from the rest of the cell contents.</CONTEXT>

<ID>6</ID>
<TITLE>FancD2</TITLE>
<NEED>Find articles about function of FancD2.</NEED>
<CONTEXT>There are many genes involved in Fanconi Anemia and the downstream pathways of FancD2 in flies. The FancD2 is monoubiquitylated and there are 2 components of the FancD2 pathway. The researcher studies the FancD2 pathway in flies.</CONTEXT>

<ID>7</ID>
<TITLE>DNA repair and oxidative stress</TITLE>
<NEED>Find correlation between DNA repair pathways and oxidative stress.</NEED>
<CONTEXT>Researcher is interested in how oxidative stress effects DNA repair.</CONTEXT>

<ID>8</ID>
<TITLE>Correlation between DNA repair pathways and skin cancer</TITLE>
<NEED>Genes and proteins (pathways) common to DNA repair, oxidative diseases, skin-carcinogenesis, and UV-carcinogenesis.</NEED>
<CONTEXT>Are there genes and mechanisms that are utilized by more than one of these fields? A relevant article mentions a gene or pathway, DNA repair, and one or more oxidative or cancerous diseases.</CONTEXT>

<ID>9</ID>
<TITLE>mutY</TITLE>
<NEED>Find articles about the function of mutY in humans.</NEED>
<CONTEXT>mutY is particularly challenging, because it is also known as hMYH. This is further complicated by the fact that myoglobin genes are also typically located in search results.</CONTEXT>

<ID>10</ID>
<TITLE>NEIL1</TITLE>
<NEED>Find articles about the role of NEIL1 in repair of DNA.</NEED>
<CONTEXT>Interested in role that NEIL1 plays in DNA repair.</CONTEXT>

<ID>11</ID>
<TITLE>Carcinogenesis and hairless mice</TITLE>
<NEED>Find articles regarding carcinogenesis induced in hairless mice.</NEED>
<CONTEXT>Researching genes and proteins (pathways) common to DNA repair, oxidative diseases, skin-carcinogenesis, and UV-carcinogenesis.</CONTEXT>

<ID>12</ID>
<TITLE>Genes regulated by Smad4</TITLE>
<NEED>Find articles describing genes that are regulated by the signal transducing molecule Smad4.</NEED>
<CONTEXT>Project is to characterize Smad4 knockout mouse in skin (specifically skin) to establish signaling network. Identify all Smad4 targets to compare gene expression patterns of the knockout mouse to the normal mouse.</CONTEXT>

<ID>13</ID>
<TITLE>Role of TGFB in angiogenesis in skin</TITLE>
<NEED>Documents regarding the role of TGFB in angiogenesis in skin with respect to homeostasis and development.</NEED>
<CONTEXT>TGFB plays a crucial role in regulating angiogenesis, a biological process that occurs during development and homeostasis, as well as during inflammatory perturbation.</CONTEXT>

<ID>14</ID>
<TITLE>Expression or Regulation of TGFB in HNSCC cancers</TITLE>
<NEED>Documents regarding TGFB expression or regulation in HNSCC cancers.</NEED>
<CONTEXT>The laboratory wants to identify components of the the TGFB signaling pathway in HNSCC, and determine new targets to study HNSCC.</CONTEXT>

<ID>15</ID>
<TITLE>ATPase and apoptosis</TITLE>
<NEED>Find information on role of ATPases in apoptosis</NEED>
<CONTEXT>The laboratory wants to know more about the role of ATPases in apoptosis.</CONTEXT>

<ID>16</ID>
<TITLE>AAA proteins</TITLE>
<NEED>How do AAA proteins mediate interaction with lipids or DNA and what is their functional impact?</NEED>
<CONTEXT>A relevant document is one that discusses protein interactions interactions involving members of the AAA protein family that can help to determine their functional importance.</CONTEXT>

<ID>17</ID>
<TITLE>DO1 antibody</TITLE>

<NEED>Determine binding affinity of anti-p53 monoclonal antibody DO1.</NEED>
<CONTEXT>One aspect of determining how an antibody works is to determine its binding affinity. A relevant document is one which discusses the binding affinity of DO1.</CONTEXT>

<ID>18</ID>
<TITLE>Gis4</TITLE>
<NEED>Properties of Gis4 with respect to cell cycle and/or metabolism.</NEED>
<CONTEXT>It is possible that Gis4 plays a role between cell cycle and yeast carbon pathways and that there is a link between cell cycle and metabolism. A relevant document is one that supports or refutes this hypothesis with regard to the properties of Gis4 in one or both processes.</CONTEXT>

<ID>19</ID>
<TITLE>Comparison of Promoters of GAL1 and SUC1</TITLE>
<NEED>What similarities and differences exist between the upstream promoter regions of GAL1 and SUC1? Are there co-repressors or co-activators? If so, are they regulated by SNF1?</NEED>
<CONTEXT>Gis4 may play a role between the cell cycle and yeast carbon pathways. SNF1 is an upstream kinase of Gis 4.</CONTEXT>

<ID>20</ID>
<TITLE>Substrate modification by ubiquitin</TITLE>
<NEED>Which biological processes are regulated by having constituent proteins modified by covalent attachment to ubiquitin or ubiquitin-like proteins?</NEED>
<CONTEXT>Ubiquitin and ubiquitin-like proteins have important roles in controlling cell division, signal transduction, embryonic development, endocytic trafficking, and the immune response.</CONTEXT>

<ID>21</ID>
<TITLE>Role of p63 and p73 in relation to DNA damage</TITLE>
<NEED>Do p63 and p73 cause cell cycle arrest or apoptosis related to DNA damage?</NEED>
<CONTEXT>DNA damage may cause cell cycle arrest or apoptosis. p63 and p73 may play a role in mediating these sequelae of DNA damage.</CONTEXT>

<ID>22</ID>
<TITLE>Relative response of p53 family members to agents causing single-stranded versus double-stranded DNA breaks</TITLE>
<NEED>Does p53 respond differently to different DNA-damaging agents? Do they respond differently to single-strand versus double-strand breaks?</NEED>
<CONTEXT>DNA damage may cause cell cycle arrest or apoptosis. p53 plays a role in mediating these sequelae of DNA damage.</CONTEXT>

<ID>23</ID>
<TITLE>Saccharomyces cerevisiae proteins involved in ubiquitin system</TITLE>
<NEED>Which Saccharomyces cerevisiae proteins are involved in the ubiquitin proteolytic pathway?</NEED>
<CONTEXT>The researcher identified a protein in another yeast species and wants to compare it to the same one in Saccharomyces cerevisiae.</CONTEXT>

<ID>24</ID>
<TITLE>Mouse peptidoglycan recognition proteins (PGRP)</TITLE>
<NEED>Find all reports describing mouse peptidoglycan recognition proteins (PGRP).</NEED>
<CONTEXT>A research group is preparing a manuscript about four poorly characterized mouse PGRP genes. Their findings include new information about gene regulation. They report longer DNA and protein sequences than those found in GenBank, and sub-cellular location discrepancies.</CONTEXT>

<ID>25</ID>
<TITLE>Cause of scleroderma</TITLE>
<NEED>Identify studies that include genome-wide scans and microarray analysis in the investigation of scleroderma.</NEED>
<CONTEXT>New information about experiments and genes involved in scleroderma.</CONTEXT>

<ID>26</ID>
<TITLE>Function of BUB2/BFA1 in the process of cytokinesis</TITLE>
<NEED>Retrieval of information regarding the role of BUB2 and BFA1 in cytokinesis in yeast.</NEED>
<CONTEXT>Information gathering for the purpose of supplementing the information from a local protocol.</CONTEXT>

<ID>27</ID>
<TITLE>Role of autophagy in apoptosis</TITLE>
<NEED>Experiments establishing positive or negative interconnection between autophagy and apoptosis.</NEED>
<CONTEXT>New information about experiments and genes involved in autophagic cell death.</CONTEXT>

<ID>28</ID>
<TITLE>Proteases that function in both apoptosis and autophagy cell death</TITLE>
<NEED>Studies that investigate similarities in morphological changes among apoptosis and autophagy processes.</NEED>
<CONTEXT>Collection of information regarding the potential relationship between apoptosis and autophagy.</CONTEXT>

<ID>29</ID>
<TITLE>Phenotypes of gyrA mutations</TITLE>
<NEED>Documents containing the sequences and phenotypes of E. coli gyrA mutations.</NEED>
<CONTEXT>The laboratory has isolated some gyrA mutations in E. coli. They want to compare their mutant gyrA with the wild-type and other mutant sequences.</CONTEXT>

<ID>30</ID>
<TITLE>Regulatory targets of the Nkx gene family members</TITLE>
<NEED>Documents identifying genes regulated by Nkx gene family members.</NEED>
<CONTEXT>The laboratory needs markers to follow Nkx family-member expression and activity.</CONTEXT>

<ID>31</ID>
<TITLE>TOR signaling in neurofibromatosis</TITLE>
<NEED>Reports that provide possible links between neurofibromatosis and TOR signaling.</NEED>
<CONTEXT>TOR is a serine-threonine kinase in a pathway involved in the control of cell growth and proliferation, and it is the target of the signaling inhibitor rapamycin.</CONTEXT>

<ID>32</ID>
<TITLE>Xenograft animal models of tumorogenesis</TITLE>
<NEED>Find reports that describe xenograft models of human cancers.</NEED>
<CONTEXT>A xenograft animal model of cancer is one in which foreign tumor tissue is grafted into animals, usually rodents, providing a means to test various compounds for their ability to slow or halt tumor growth.</CONTEXT>

<ID>33</ID>
<TITLE>Mice, mutant strains, and Histoplasmosis</TITLE>
<NEED>Identify research on mutant mouse strains and factors which increase susceptibility to infection by Histoplasma capsulatum.</NEED>

<CONTEXT>The ultimate goal of this initial research study, is to identify mouse genes that will influence the outcome of blood borne pathogen infections.</CONTEXT>

<ID>34</ID>
<TITLE>Gene products of Cryptococcus important to fungal survival</TITLE>
<NEED>Articles reporting experiments allowing annotation of gene products of Cryptococcus.</NEED>
<CONTEXT>Information needed to contribute to the development of a standardized annotated database of Cryptococcus neoformans genome.</CONTEXT>

<ID>35</ID>
<TITLE>WD40 repeat-containing proteins</TITLE>
<NEED>What is the function of proteins containing WD40 repeats?</NEED>
<CONTEXT>Need to understand the variety of functions that involve this domain.</CONTEXT>

<ID>36</ID>
<TITLE>RAB3A</TITLE>
<NEED>Background information on RAB3A.</NEED>
<CONTEXT>Further information about a gene is needed after it is identified through a gene expression profile. The genes are related to synaptic plasticity in learning and memory.</CONTEXT>

<ID>37</ID>
<TITLE>PAM</TITLE>
<NEED>What research is being done on peptide amidating enzyme, PAM?</NEED>
<CONTEXT>Need to put specific PAM research in the context of other researchers work.</CONTEXT>

<ID>38</ID>
<TITLE>Risk factors for stroke</TITLE>
<NEED>Information concerning genetic loci that are associated with increased risk of stroke, such as apolipoprotein E4 or factor V mutations.</NEED>
<CONTEXT>Candidate gene testing within a large Scottish case-control study of genetic risk factors for stroke. Future research includes investigations into other ethnically distinct populations.</CONTEXT>

<ID>39</ID>
<TITLE>Hypertension</TITLE>
<NEED>Identify genes as potential genetic risk factors candidates for causing hypertension.</NEED>
<CONTEXT>A relevant document is one which discusses genes that could be considered as candidates to test in a randomized controlled trial which studies the genetic risk factors for stroke.</CONTEXT>

<ID>40</ID>
<TITLE>Antigens expressed by lung epithelial cells</TITLE>
<NEED>To identify the antigens expressed by lung epithelial cells and the antibodies available.</NEED>
<CONTEXT>Information gathering to design assays to determine the nature of donor cells in tissues of chimaeric animals.</CONTEXT>

<ID>41</ID>
<TITLE>Mutations in the Cystic Fibrosis conductance regulator gene</TITLE>
<NEED>What phenotypes have been described resulting from mutations in the Cystic Fibrosis conductance regulator gene?</NEED>
<CONTEXT>Comparing protein mutations detected utilizing mass spectrometry.</CONTEXT>

<ID>42</ID>
<TITLE>Genes altered by chromosome translocations</TITLE>
<NEED>What genes show altered behavior due to chromosomal rearrangements?</NEED>
<CONTEXT>Information is required on the disruption of functions from genomic DNA rearrangements.</CONTEXT>

<ID>43</ID>
<TITLE>Sleeping Beauty</TITLE>
<NEED>Studies of Sleeping Beauty transposons.</NEED>
<CONTEXT>A relevant document is one that discusses studies on Sleeping Beauty. Interviewee's group studies a related element and want to know what others are doing in a similar field.</CONTEXT>

<ID>44</ID>
<TITLE>Proteins involved in the nerve growth factor pathway</TITLE>
<NEED>Create a list of all the nerve growth factor pathway proteins.</NEED>
<CONTEXT>Need to identify genes that are most likely to be involved in the nerve growth factor pathway.</CONTEXT>

<ID>45</ID>
<TITLE>Mental Health Wellness-1</TITLE>
<NEED>What genetic loci, such as Mental Health Wellness 1 (MWH1) are implicated in mental health?</NEED>
<CONTEXT>Want to identify genes involved in mental disorders.</CONTEXT>

<ID>46</ID>
<TITLE>RSK2</TITLE>
<NEED>What human biological processes is RSK2 known to be involved in?</NEED>
<CONTEXT>After being identified via microarrays, the biological processes the genes are involved in needs to be discovered.</CONTEXT>

<ID>47</ID>
<TITLE>Human gene BCL-2 antagonists and inhibitors</TITLE>
<NEED>Research the human gene BCL-2 to determine if there are antagonists and inhibitors inside of a cell.</NEED>
<CONTEXT>Early research goals included learning more about BCL2-interacting molecules, which facilitated identifying new inhibitors during preliminary testing.</CONTEXT>

<ID>48</ID>
<TITLE>Human homologues of C. elegans UNC genes</TITLE>
<NEED>What is the focus of studies involving the members of the human UNC gene family?</NEED>
<CONTEXT>The interviewee wished to determine the interests and focus of a fellow scientist that was investigating similar topics to their own.</CONTEXT>

<ID>49</ID>
<TITLE>Glyphosate tolerance gene sequence</TITLE>
<NEED>Find reports and glyphosate tolerance gene sequences in the literature.</NEED>
<CONTEXT>A DNA sequence isolated in the laboratory is often sequenced only partially, until enough sequence is generated to identify the gene. In these situations, the rest of the sequence is inferred from matching clones in the public domain. When there is difficulty in the laboratory manipulating the DNA segment using sequence-dependent methods, the laboratory isolate must be re-examined.</CONTEXT>

<ID>50</ID>
<TITLE>Low temperature protein expression in E. coli</TITLE>

<NEED>Find research on improving protein expressions at low temperature in Escherichia coli bacteria.</NEED>
<CONTEXT>The researcher is not satisfied with the yield of expressing a protein in E. coli when grown at low temperature and is searching for a better solution. The researcher is willing to try a different organism and/or method.</CONTEXT>

Appendix III – stop words

Source: http://www.princeton.edu/~biolib/instruct/MedSW.html

Stopwords for Medline				
a	due	mg	somewhat	than
accordingly	during	might	regardless	that
affected	do	knowledge	predominantly	the
affecting	does	largely	present	their
affects	done	like	previously	theirs
after	each	made	primarily	them
again	effect	mainly	probably	then
against	either	make	prompt	there
all	else	many	promptly	therefore
almost	enough	may	quickly	these
already	especially	ml	quite	they
also	et-al	more	rather	this
although	etc	most	readily	those
always	ever	mostly	really	though
among	every	much	recently	through
an	following	mug	refs	throughout
and	for	must	relatively	to
another	found	nearly	respectively	too
any	from	necessarily	resulted	toward
anyone	further	neither	resulting	under
apparently	gave	next	results	until
are	gets	no	said	upon
arise	give	none	same	use
as	given	nor	seem	used
aside	giving	normally	seen	usefully
at	gone	nos	several	usefully
away	got	not	should	usefulness
be	had	noted	show	using
because	has	now	showed	usually
become	hardly	obtain	shown	various
becomes	have	obtained	shows	was
been	having	of	significantly	were
before	her	often	similar	what
being	how	on	similarly	when
between	however	only	since	where
biol	if	or	slightly	whether
both	immediately	other	so	which
briefly	importance	ought	some	while
but	important	our	sometime	who

by	in	out	somewhat	whose
came	into	owing	soon	why
can	is	particularly	specifically	widely
cannot	it	past	state	will
certain	its	perhaps	states	with
certainly	itself	please	strongly	within
chem	just	poorly	substantially	without
copyright	keep	possible	successfully	would
could	kept	possibly	such	yet
did	kg	potentially	sufficiently	
different	km			

Appendix IV – punctuation signs

The following punctuation signs will be replaced by spaces:

.
!
?
,
:
;
>
<
--
(
)
\
/
[
]

after replacing those punctuation sighs all characters not being [a-z] [0-9] or a whitespace will be deleted.

Appendix V – Recall values

Topic	Recall										
	run1	*run2*	*run3*	*run4*	*run5*	*run6*	*run7*	*run8*	*run9*	*run10*	*run11*
1	0,0253	0,0506	0,0506	0,3671	0,4177	0,4177	0,4177	0,4177	0.0127	0.4304	0.4304
2	0,0099	0,0198	0,0198	0,1980	0,3168	0,3663	0,3663	0,3663	0.0000	0.2277	0.0792
3	0,0000	0,0000	0,0000	0,1271	0,0884	0,0884	0,0884	0,0884	0.0000	0.0884	0.1878
4	0,1000	0,1000	0,1000	0,1000	0,3667	0,3667	0,3667	0,3667	0.1000	0.3000	0.1333
5	0,0000	0,0000	0,0000	0,0000	0,0833	0,0833	0,0833	0,0833	0.0000	0.0833	0.1250
6	0,0000	0,0000	0,0000	0,3936	0,3936	0,3936	0,3936	0,3936	0.0000	0.3936	0.3936
7	0,4348	0,4348	0,4348	0,1652	0,3652	0,4348	0,4348	0,4348	0.6087	0.7391	0.7739
8	0,0435	0,2981	0,2981	0,2422	0,3540	0,3727	0,3727	0,3727	0.0000	0.3727	0.3975
9	0,3304	0,3304	0,3304	1,0000	0,9478	0,9478	0,9478	0,9478	0.0000	0.9478	1.0000
10	0,5000	0,5000	0,5000	0,7500	0,7500	0,7500	0,7500	0,7500	0.0000	0.7500	0.7500
11	0,2072	0,4865	0,4865	0,4865	0,3243	0,3423	0,3423	0,3423	0.0000	0.3423	0.5315
12	0,1250	0,1250	0,1250	0,1250	0,4023	0,4023	0,4023	0,4023	0.0000	0.4023	0.5117
13	0,0417	0,3750	0,3750	0,3750	0,2500	0,3333	0,3333	0,3333	0.0000	0.3333	0.3750
14	0,0000	0,0000	0,0000	0,0000	0,0476	0,0476	0,0476	0,0476	0.0000	0.0476	0.0952
15	0,5889	0,5889	0,5889	0,5889	0,4000	0,4000	0,4000	0,4000	0.5333	0.7444	0.8667
16	0,6395	0,6395	0,6395	0,6395	0,7415	0,7415	0,7415	0,7415	0.0000	0.7415	0.8571
17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0.0000	0.0000
18	0,0000	0,0000	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0.0000	1,0000	1,0000
19	0,0000	0,0000	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0.0000	1,0000	1,0000
20	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0,0000	0,0000
21	0,1000	0,1000	0,1000	0,3375	0,3375	0,3375	0,3375	0,3375	0.0000	0.3375	0.3375
22	0,2857	0,2857	0,2857	0,0476	0,6333	0,6476	0,6476	0,6476	0.2000	0.5619	0.5619
23	0,6076	0,6076	0,6076	0,6076	0,5063	0,5063	0,5063	0,5063	0.5823	0.7595	0.8797
24	0,7692	0,8846	0,8846	0,8846	0,9615	0,9615	0,9615	0,9615	0.0000	0.9615	0.9615
25	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0.0000	0.0000
26	0,1064	0,1064	0,1064	0,1064	0,0638	0,0638	0,0638	0,0638	0.0000	0.0638	0.1064
27	0,8276	0,8276	0,8276	0,8276	0,6207	0,6207	0,6207	0,6207	0.0000	0.6207	0.8276
28	1,0000	1,0000	1,0000	1,0000	0,6923	0,6923	0,6923	0,6923	0.0000	0.6923	1.0000
29	0,5581	0,5581	0,5581	0,5581	0,7907	0,7907	0,7907	0,7907	0.0000	0.7907	0.9070
30	0,1879	0,1879	0,1879	0,1879	0,1879	0,1879	0,1879	0,1879	0.0000	0.1879	0.1879
31	0,0000	0,0000	0,0000	0,2971	0,3478	0,3478	0,3478	0,3478	0.0000	0.3478	0.3478
32	0,1935	0,2500	0,2500	0,2540	0,6048	0,6048	0,6048	0,6048	0.0000	0.4294	0.4476
33	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0.0000	0.0000
34	0,4194	0,4516	0,4516	0,4516	0,4194	0,4194	0,4194	0,4194	0.6129	0.8065	0.8710
35	0,2399	0,2399	0,2399	0,6531	0,6531	0,6531	0,6531	0,6531	0.0000	0.6531	0.6531
36	0,8150	0,8150	0,8150	0,8150	0,8150	0,8150	0,8150	0,8150	0.2402	0.8386	0.8386
37	0,6242	0,6980	0,6980	0,6980	0,8725	0,8725	0,8725	0,8725	0.0067	0.8725	0.8993
38	0,0662	0,0662	0,0662	0,0520	0,1513	0,1513	0,1513	0,1513	0.0000	0.0757	0.0284
39	0,0252	0,0631	0,0631	0,0662	0,0536	0,0820	0,5962	0,5962	0.0063	0.0442	0.0284
40	0,0830	0,0830	0,0830	0,1011	0,3394	0,3394	0,3394	0,3394	0.1336	0.4368	0.6282
41	0,2354	0,3179	0,3179	0,3196	0,5893	0,5962	0,5962	0,5962	0.4192	0.5515	0.8557
42	0,1478	0,1478	0,1478	0,1248	0,3960	0,3960	0,3960	0,3960	0.0588	0.2640	0.1693
43	0,1282	0,1282	0,1282	0,1282	0,1128	0,1179	0,1179	0,1179	0.0000	0.1179	0.1282
44	0,2496	0,2496	0,2496	0,1356	0,5840	0,5840	0,5840	0,5840	0.0570	0.1926	0.1772
45	0,0385	0,0385	0,0385	0,0385	0,0385	0,0385	0,0385	0,0385	0.0000	0.0385	0.0641
46	0,0558	0,0558	0,0558	0,4416	0,2030	0,2589	0,2589	0,2589	0.2234	0.4162	0.0457

47	0,0082	0,0082	0,0082	0,1397	0,1890	0,3890	0,3890	0,3890	0.0767	0.2000	0.1452
48	0,0194	0,0194	0,0194	0,0323	0,0323	0,0323	0,0323	0,0323	0.0000	0.0323	0.0323
49	0,1781	0,1781	0,1781	0,1781	0,1507	0,1507	0,3151	0,3151	0.0000	0.3151	0.3836
50	0,2152	0,2152	0,2152	0,2152	0,2185	0,2185	0,2351	0,2351	0.0000	0.2351	0.3311
all	0,2674	0,2845	0,2845	0,3606	0,4216	0,4333	0,4371	0,4371	0.1613	0.4331	0.4657

Appendix VI – MAP values

Topic	Map										
	run1	*run2*	*run3*	*run4*	*run5*	*run6*	*run7*	*run8*	*run9*	*run10*	*run11*
1	0,0253	0,0140	0,0140	0,0988	0,0565	0,2598	0,3357	0,3276	0.0005	0.2862	0.2225
2	0,0050	0,0116	0,0116	0,0052	0,0027	0,0127	0,0126	0,0191	0.0000	0.0100	0.0006
3	0,0000	0,0000	0,0000	0,0075	0,0055	0,0112	0,0143	0,0136	0.0000	0.0143	0.0157
4	0,0005	0,0005	0,0005	0,0005	0,0015	0,0042	0,0039	0,0041	0.0009	0.0032	0.0008
5	0,0000	0,0000	0,0000	0,0000	0,0008	0,0024	0,0065	0,0422	0.0000	0.0064	0.0060
6	0,0000	0,0000	0,0000	0,3350	0,3350	0,3668	0,3668	0,3628	0.0000	0.3668	0.3668
7	0,0480	0,0480	0,0480	0,0036	0,0054	0,0429	0,1307	0,1457	0.1147	0.1941	0.2162
8	0,0081	0,0417	0,0417	0,0127	0,0197	0,0465	0,0448	0,0439	0.0000	0.0443	0.0423
9	0,2331	0,2331	0,2331	0,5529	0,5287	0,7676	0,7676	0,7791	0.0000	0.7676	0.8270
10	0,5000	0,0917	0,0917	0,0956	0,0938	0,0792	0,0870	0,0725	0.0000	0.0870	0.0834
11	0,1528	0,3892	0,3892	0,3892	0,2924	0,2883	0,2883	0,2836	0.0000	0.2838	0.4146
12	0,0611	0,0611	0,0611	0,0611	0,1679	0,3037	0,3037	0,3034	0.0000	0.3037	0.3956
13	0,0208	0,0075	0,0075	0,0075	0,0037	0,0267	0,0721	0,0722	0.0000	0.0611	0.0141
14	0,0000	0,0000	0,0000	0,0000	0,0119	0,0476	0,0476	0,0476	0.0000	0.0476	0.0231
15	0,1742	0,1742	0,1742	0,1742	0,1222	0,2852	0,2852	0,2916	0.2357	0.3811	0.4677
16	0,1851	0,1851	0,1851	0,1851	0,1988	0,1988	0,1988	0,1988	0.0000	0.1988	0.1936
17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0.0000	0.0000
18	0,0000	0,0000	0,0000	1,0000	1,0000	1,0000	1,0000	1,0000	0.0000	1,0000	1,0000
19	0,0000	0,0000	0,0000	0,0090	0,0091	0,0500	0,0714	0,0500	0.0000	0.0714	0.0385
20	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0,0000	0,0000
21	0,0501	0,0501	0,0501	0,0921	0,0921	0,0905	0,1057	0,1058	0.0000	0.1057	0.1020
22	0,0227	0,0227	0,0227	0,0004	0,0477	0,0760	0,1389	0,1328	0.0161	0.1287	0.1150
23	0,2880	0,2880	0,2880	0,2880	0,2250	0,2682	0,2682	0,2641	0.1335	0.3501	0.2006
24	0,6833	0,3791	0,3791	0,3791	0,3544	0,7937	0,8128	0,8257	0.0000	0.8128	0.7185
25	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0.0000	0.0000
26	0,1064	0,1064	0,1064	0,1064	0,0638	0,0638	0,0638	0,0638	0.0000	0.0638	0.1064
27	0,3060	0,3060	0,3060	0,3060	0,3387	0,4539	0,4539	0,4177	0.0000	0.4539	0.4732
28	0,1952	0,1952	0,1952	0,1952	0,1814	0,2045	0,2045	0,1894	0.0000	0.2045	0.2617
29	0,0509	0,0509	0,0509	0,0509	0,0822	0,1105	0,1105	0,0996	0.0000	0.1105	0.1220
30	0,1438	0,1438	0,1438	0,1438	0,1438	0,1427	0,1427	0,1452	0.0000	0.1427	0.1417
31	0,0000	0,0000	0,0000	0,0171	0,0178	0,0328	0,0328	0,0420	0.0000	0.0328	0.0336
32	0,0642	0,0645	0,0645	0,0614	0,0870	0,1723	0,1723	0,1841	0.0000	0.1392	0.1148
33	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0.0000	0.0000
34	0,0089	0,0111	0,0111	0,0111	0,0089	0,0199	0,0199	0,0431	0.0182	0.0374	0.0597
35	0,2173	0,2173	0,2173	0,5597	0,5597	0,6050	0,6050	0,6014	0.0000	0.6050	0.6073
36	0,6996	0,6519	0,6519	0,6519	0,5536	0,7027	0,7602	0,7558	0.1989	0.7853	0.5940
37	0,1600	0,1762	0,1762	0,1762	0,1751	0,4214	0,3206	0,2782	0.0004	0.3194	0.4234
38	0,0069	0,0069	0,0069	0,0046	0,0033	0,0047	0,0047	0,0051	0.0000	0.0023	0.0004
39	0,0015	0,0019	0,0019	0,0019	0,0003	0,0011	0,2341	0,0012	0.0000	0.0006	0.0002
40	0,0150	0,0150	0,0150	0,0217	0,0553	0,0818	0,0862	0,1005	0.0472	0.1225	0.2051
41	0,1367	0,1100	0,1100	0,1106	0,1914	0,2357	0,2341	0,2562	0.2016	0.2520	0.3377
42	0,0272	0,0272	0,0272	0,0158	0,0453	0,0852	0,0852	0,0888	0.0046	0.0650	0.0260
43	0,0986	0,0986	0,0986	0,0092	0,0077	0,0055	0,0049	0,0050	0.0000	0.0049	0.0062
44	0,0822	0,0822	0,0822	0,0165	0,0666	0,0666	0,0892	0,0880	0.0055	0.0313	0.0257
45	0,0280	0,0280	0,0280	0,0280	0,0280	0,0339	0,0310	0,0310	0.0000	0.0310	0.0156
46	0,0464	0,0158	0,0158	0,0252	0,0050	0,1150	0,1789	0,1730	0.0893	0.2714	0.0056

47	0,0026	0,0026	0,0026	0,0097	0,0051	0,0396	0,0474	0,0482	0.0039	0.0327	0.0129
48	0,0075	0,0075	0,0075	0,0020	0,0018	0,0015	0,0015	0,0013	0.0000	0.0015	0.0014
49	0,1395	0,1395	0,1395	0,1395	0,1063	0,1063	0,2388	0,2402	0.0000	0.2388	0.2799
50	0,0346	0,0346	0,0346	0,0346	0,0336	0,0336	0,0483	0,0455	0.0000	0.0483	0.0697
all	0,1199	0,1020	0,1020	0,1361	0,1348	0,1864	0,1978	0,1977	0.0446	0.1984	0.1956

Appendix VII - P10 values

Topic	P10										
	run1	*run2*	*run3*	*run4*	*run5*	*run6*	*run7*	*run8*	*run9*	*run10*	*run11*
1	0,2000	0,3000	0,3000	0,0000	0,0000	0,8000	1,0000	0,9000	0.0000	1,0000	0.7000
2	0,1000	0,2000	0,2000	0,0000	0,0000	0,0000	0,0000	0,2000	0.0000	0.0000	0.0000
3	0,0000	0,0000	0,0000	0,1000	0,1000	0,0000	0,1000	0,2000	0.0000	0.1000	0.0000
4	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0.0000	0.0000
5	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,1000	0,1000	0.0000	0.1000	0.1000
6	0,0000	0,0000	0,0000	0,9000	0,9000	0,9000	0,9000	0,9000	0.0000	0.9000	0.9000
7	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,5000	0,4000	0.1000	0.6000	0.5000
8	0,2000	0,0000	0,0000	0,0000	0,0000	0,1000	0,1000	0,0000	0.0000	0.1000	0.1000
9	0,5000	0,5000	0,5000	0,7000	0,7000	1,0000	1,0000	1,0000	0.0000	1,0000	1,0000
10	0,2000	0,2000	0,2000	0,0000	0,0000	0,1000	0,1000	0,0000	0.0000	0.1000	0.0000
11	0,8000	0,7000	0,7000	0,7000	0,9000	0,8000	0,8000	0,8000	0.0000	0.8000	0.6000
12	0,4000	0,4000	0,4000	0,4000	0,3000	1,0000	1,0000	1,0000	0.0000	1,0000	1,0000
13	0,1000	0,0000	0,0000	0,0000	0,0000	0,0000	0,3000	0,3000	0.0000	0.3000	0.0000
14	0,0000	0,0000	0,0000	0,0000	0,1000	0,1000	0,1000	0,1000	0.0000	0.1000	0.2000
15	0,4000	0,4000	0,4000	0,4000	0,3000	0,9000	0,9000	0,9000	0.9000	0.8000	0.7000
16	0,3000	0,3000	0,3000	0,3000	0,3000	0,3000	0,3000	0,3000	0.0000	0.3000	0.3000
17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0.0000	0.0000
18	0,0000	0,0000	0,0000	0,1000	0,1000	0,1000	0,1000	0,1000	0.0000	0.1000	0.1000
19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0.0000	0.0000
20	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0,0000	0,0000
21	0,3000	0,3000	0,3000	0,2000	0,2000	0,2000	0,4000	0,3000	0.0000	0.4000	0.2000
22	0,2000	0,2000	0,2000	0,0000	0,2000	0,2000	0,6000	0,4000	0.1000	0.5000	0.3000
23	0,5000	0,5000	0,5000	0,5000	0,4000	0,5000	0,5000	0,5000	0.2000	0.5000	0.2000
24	0,9000	0,2000	0,2000	0,2000	0,2000	0,9000	0,9000	1,0000	0.0000	0.9000	0.8000
25	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0.0000	0.5000
26	0,0000	0,0000	0,0000	0,5000	0,3	0,3	0,3	0,3	0.0000	0.3000	0.6000
27	0,3000	0,3000	0,3000	0,3000	0,5000	0,8000	0,8000	0,7000	0.0000	0.8000	0.6000
28	0,0000	0,0000	0,0000	0,0000	0,2000	0,3000	0,3000	0,1000	0.0000	0.3000	0.2000
29	0,0000	0,0000	0,0000	0,0000	0,1000	0,1000	0,1000	0,2000	0.0000	0.1000	0.2000
30	0,8000	0,8000	0,8000	0,8000	0,8000	0,8000	0,8000	0,9000	0.0000	0.8000	0.8000
31	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0.0000	0.0000
32	0,5000	0,1000	0,1000	0,1000	0,0000	0,5000	0,5000	0,8000	0.0000	0.5000	0.7000
33	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0.0000	0.0000
34	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,2000	0.0000	0.0000	0.1000
35	0,9000	0,9000	0,9000	1,0000	1,0000	1,0000	1,0000	1,0000	0.0000	1,0000	1,0000
36	1,0000	0,8000	0,8000	0,8000	0,9000	0,9000	1,0000	1,0000	0.8000	1,0000	0,9000
37	0,4000	0,4000	0,4000	0,4000	0,4000	0,7000	0,6000	0,6000	0.0000	0.6000	0.6000
38	0,2000	0,2000	0,2000	0,1000	0,1000	0,0000	0,0000	0,0000	0.0000	0.0000	0.0000
39	0,1000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0,0000	0,0000
40	0,0000	0,0000	0,0000	0,1000	0,2000	0,5000	0,5000	0,6000	0.2000	0.5000	0.6000
41	0,4000	0,2000	0,2000	0,2000	0,1000	0,4000	0,5000	0,4000	0.6000	0.5000	0.8000
42	0,3000	0,3000	0,3000	0,2000	0,2000	0,6000	0,6000	0,3000	0.1000	0.4000	0.2000
43	0,7000	0,7000	0,7000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0.0000	0.0000
44	0,0000	0,0000	0,0000	0,0000	0,1000	0,1000	0,0000	0,0000	0.2000	0.1000	0.4000

45	0,6000	0,6000	0,6000	0,6000	0,6000	0,6000	0,6000	0,6000	0.0000	0.6000	0.2000
46	0,7000	0,3000	0,3000	0,3000	0,0000	0,7000	0,9000	1,0000	0.5000	0.9000	0.2000
47	0,1000	0,1000	0,1000	0,2000	0,0000	0,4000	0,4000	0,6000	0.1000	0.4000	0.2000
48	0,3000	0,3000	0,3000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0.0000	0.0000
49	0,7000	0,7000	0,7000	0,7000	0,7000	0,7000	0,6000	0,6000	0.0000	0.6000	0.6000
50	0,1000	0,1000	0,1000	0,1000	0,1000	0,1000	0,4000	0,3000	0.0000	0.4000	0.1000
all	0,3262	0,2591	0,2591	0,2319	0,2340	0,3702	0,4170	0,4170	0.1583	0.4042	0.3458

Appendix VIII - P100 values

Topic	P100										
	run1	*run2*	*run3*	*run4*	*run5*	*run6*	*run7*	*run8*	*run9*	*run10*	*run11*
1	0,0200	0,0400	0,0400	0,2900	0,1300	0,2600	0,2800	0,2800	0.0100	0.2900	0.2400
2	0,0100	0,0200	0,0200	0,0000	0,0000	0,0600	0,0600	0,0800	0.0000	0.0600	0.0000
3	0,0000	0,0000	0,0000	0,0400	0,0500	0,1300	0,1300	0,1100	0.0000	0.1300	0.1100
4	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0.0000	0.0000
5	0,0000	0,0000	0,0000	0,0000	0,0000	0,0100	0,0100	0,0100	0.0000	0.0100	0.0100
6	0,0000	0,0000	0,0000	0,3700	0,3700	0,3700	0,3700	0,3700	0.0000	0.3700	0.3700
7	0,1000	0,1000	0,1000	0,0000	0,0000	0,0600	0,2600	0,2900	0.2000	0.2900	0.2900
8	0,0700	0,0900	0,0900	0,0300	0,0400	0,1300	0,1400	0,1200	0.0000	0.1400	0.1100
9	0,3800	0,3800	0,3800	0,5000	0,5200	0,7600	0,7600	0,7500	0.0000	0.7600	0.8100
10	0,0200	0,0200	0,0200	0,0300	0,0300	0,0300	0,0300	0,0300	0.0000	0.0300	0.0300
11	0,2300	0,5400	0,5400	0,5400	0,3600	0,3800	0,3800	0,3800	0.0000	0.3800	0.5900
12	0,3200	0,3200	0,3200	0,3200	0,4200	1,0000	0,6600	0,7000	0.0000	0.6600	0.7500
13	0,0100	0,0100	0,0100	0,0100	0,0100	0,0500	0,0600	0,0500	0.0000	0.0500	0.0400
14	0,0000	0,0000	0,0000	0,0000	0,0100	0,0100	0,0100	0,0100	0.0000	0.0100	0.0200
15	0,2600	0,2600	0,2600	0,2600	0,2600	0,3500	0,3500	0,3500	0.2800	0.3400	0.4600
16	0,2000	0,2000	0,2000	0,2000	0,1900	0,1900	0,1900	0,1900	0.0000	0.1900	0.2200
17	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0.0000	0.0000
18	0,0000	0,0000	0,0000	0,0100	0,0100	0,0100	0,0100	0,0100	0.0000	0.0100	0.0100
19	0,0000	0,0000	0,0000	0,0000	0,0000	0,0100	0,0100	0,0100	0.0000	0.0100	0.0100
20	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0,0000	0,0000
21	0,0800	0,0800	0,0800	0,2700	0,2700	0,2700	0,2700	0,2700	0.0000	0.2700	0.2700
22	0,0300	0,0300	0,0300	0,0000	0,0700	0,1500	0,2800	0,3200	0.0500	0.2900	0.2800
23	0,4900	0,4900	0,4900	0,4900	0,4300	0,5400	0,5400	0,5400	0.2700	0.5800	0.2300
24	0,2000	0,2300	0,2300	0,2300	0,2500	0,2500	0,2500	0,2500	0.0000	0.2500	0.2400
25	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0.0000	0.0000
26	0,0000	0,0000	0,0000	0,0500	0,0300	0,0300	0,0300	0,0300	0.0000	0.0300	0.0500
27	0,2400	0,2400	0,2400	0,2400	0,1800	0,1800	0,1800	0,1800	0.0000	0.1800	0.2400
28	0,1300	0,1300	0,1300	0,1300	0,0900	0,0900	0,0900	0,0900	0.0000	0.0900	0.1300
29	0,1100	0,1100	0,1100	0,1100	0,0700	0,1400	0,1400	0,1200	0.0000	0.1400	0.1300
30	0,3100	0,3100	0,3100	0,3100	0,3100	0,3100	0,3100	0,3100	0.0000	0.3100	0.3100
31	0,0000	0,0000	0,0000	0,0300	0,0300	0,1200	0,1200	0,1800	0.0000	0.1200	0.1200
32	0,2800	0,2500	0,2500	0,2400	0,1700	0,4200	0,4200	0,5000	0.0000	0.4200	0.4000
33	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0.0000	0.0000	0.0000
34	0,0100	0,0200	0,0200	0,0200	0,0100	0,0600	0,0600	0,0600	0.0200	0.0400	0.0500
35	0,6500	0,6500	0,6500	0,8500	0,8500	0,9200	0,9200	0,9200	0.0000	0.9200	0.9200
36	0,8800	0,8500	0,8500	0,8500	0,7200	0,9900	0,9800	0,9600	0.6100	1,0000	0.8800
37	0,2500	0,2500	0,2500	0,2600	0,2300	0,5000	0,4000	0,3500	0.0100	0.4100	0.5700
38	0,0800	0,0800	0,0800	0,0600	0,0200	0,0000	0,0000	0,0000	0.0000	0.0000	0.0100
39	0,0300	0,0400	0,0400	0,0400	0,0000	0,0100	0,0000	0,0000	0.0000	0,0000	0.0000
40	0,2100	0,2100	0,2100	0,2200	0,1500	0,2300	0,3000	0,3200	0.3400	0.3000	0.3700
41	0,6200	0,3300	0,3300	0,3300	0,3300	0,4300	0,4400	0,5000	0.4800	0.5200	0.5100
42	0,1900	0,1900	0,1900	0,1800	0,1600	0,2800	0,2800	0,3200	0.0400	0.3000	0.1900
43	0,2500	0,2500	0,2500	0,0300	0,0300	0,0000	0,0000	0,0100	0.0000	0.0000	0.0000
44	0,3300	0,3300	0,3300	0,1300	0,0900	0,0900	0,2000	0,2000	0.0700	0.2000	0.1400
45	0,0600	0,0600	0,0600	0,0600	0,0600	0,0600	0,0600	0,0600	0.0000	0.0600	0.1000
46	0,1100	0,0300	0,0300	0,0800	0,0400	0,3200	0,4400	0,4400	0.3400	0.5100	0.0500

47	0,0300	0,0300	0,0300	0,0300	0,0200	0,1500	0,2000	0,2000	0.0300	0.2000	0.1100
48	0,0300	0,0300	0,0300	0,0400	0,0400	0,0400	0,0400	0,0400	0.0000	0.0400	0.0400
49	0,1300	0,1300	0,1300	0,1300	0,1100	0,1100	0,2300	0,2300	0.0000	0.2300	0.2800
50	0,0650	0,0650	0,0650	0,1700	0,1400	0,1400	0,2100	0,2100	0.0000	0.2100	0.2800
all	0,1802	0,1714	0,1714	0,1740	0,1553	0,2191	0,2302	0,2415	0.1146	0.2365	0.2285

Appendix IX - Query terms

Topic	Terms			
1	ferroportin-1	human		
2	protocol	generate	transgenic	mouse
3	time course	gene expression	murine	kidney
4	mouse	gene	kidney	
5	Subcellular fractionation	nuclei		
6	FancD2	fly		
7	DNA repair	oxidative stress		
8	DNA repair	nonmelanoma		
9	mutY	human		
10	NEIL1	DNA repair		
11	Carcinogenesis	hairless mouse		
12	Smad4	regulate		
13	TGFB	angiogenesis		
14	TGFB	HNSCC cancers		
15	ATPase	apoptosis		
16	AAA protein			
17	DO1	binding affinity		
18	Gis4	cell		
19	GAL1	SUC1		
20	ubiquitin	constituent proteins	covalent attachment	
21	p63	p73	DNA damage	
22	p53	dna damage		
23	Saccharomyces cerevisiae	ubiquitin		
24	PGRP			
25	scleroderma	scans	microarray	
26	BUB2	BFA1	cytokinesis	
27	Autophagy	apoptosis		
28	Autophagy	apoptosis		
29	gyrA	mutation		
30	Nkx			
31	neurofibromatosis	TOR		
32	xenograft	cancer	human	
33	Mutant	mouse	Histoplasmosis	
34	Cryptococcus	fungal		
35	WD40 repeat			
36	RAB3A			
37	PAM			
38	Risk factor	stroke		
39	Risk factor	Hypertension		
40	Antigens	lung epithelial cells		
41	Mutation	cf		
42	chromosome	translocations		
43	sbt			
44	nerve growth factor pathway			
45	Mental health	loci		
46	RSK2	human		
47	BCL-2	antagonists	inhibitors	
48	UNC	Human		
48	glyphosate tolerance			
50	Escherichia coli	low temperature		

Appendix X - Updated Query terms

topic	terms		
1	ferroportin-1		
2	generate	transgenic	mouse
3	gene expression	murine	kidney
4	mouse	gene	kidney
5	Subcellular fractionation	nuclei	
6	FancD2		
7	DNA repair	oxidative stress	
8	DNA	nonmelanoma	
9	mutY		
10	NEIL1	DNA	
11	Carcinogenesis	hairless mouse	
12	Smad4	regulate	
13	TGFB	angiogenesis	
14	TGFB	HNSCC cancers	
15	ATPase	apoptosis	
16	AAA protein		
17	DO1	binding affinity	
18	Gis4		
19	GAL1		
20	ubiquitin	constituent proteins	covalent attachment
21	p63	p73	
22	p53	dna damage	
23	Saccharomyces cerevisiae	ubiquitin	
24	PGRP		
25	scleroderma	scans	microarray
26	BUB2	BFA1	cytokinesis
27	Autophagy	apoptosis	
28	Autophagy	apoptosis	
29	gyrA	mutation	
30	Nkx		
31	TOR		
32	xenograft	cancer	
33	Mutant	Histoplasmosis	
34	Cryptococcus	fungal	
35	WD40		
36	RAB3A		
37	PAM		
38	Risk	stroke	
39	Risk	Hypertension	
40	Antigens	lung epithelial	
41	Mutation	cf	
42	chromosome	translocation	
43	sbt		
44	nerve growth factor		
45	Mental health	loci	
46	RSK2		
47	BCL-2		
48	UNC		
49	glyphosate tolerance		
50	Escherichia coli	low temperature	

APPENDIX XI - TREC results for every topic

	Prec(10)			Prec(100)			Average Precision		
Topic	***Best***	***Median***	***Worst***	***Best***	***Median***	***Worst***	***Best***	***Median***	***Worst***
1	1,00	0,90	0,00	0,49	0,38	0,06	0,52	0,45	0,02
2	0,60	0,40	0,00	0,18	0,13	0,00	0,11	0,07	0,00
3	0,90	0,30	0,00	0,48	0,20	0,00	0,24	0,05	0,00
4	0,20	0,00	0,00	0,11	0,02	0,00	0,06	0,01	0,00
5	0,30	0,10	0,00	0,05	0,01	0,00	0,10	0,01	0,00
6	1,00	0,90	0,00	0,52	0,39	0,02	0,57	0,37	0,01
7	0,80	0,50	0,00	0,47	0,23	0,00	0,40	0,16	0,00
8	0,60	0,40	0,00	0,31	0,16	0,00	0,18	0,05	0,00
9	1,00	1,00	0,10	0,84	0,61	0,11	0,86	0,48	0,13
10	0,40	0,30	0,00	0,04	0,03	0,00	1,00	0,68	0,00
11	1,00	0,70	0,00	0,67	0,50	0,00	0,66	0,42	0,00
12	1,00	0,90	0,00	0,87	0,71	0,00	0,76	0,52	0,00
13	0,30	0,10	0,00	0,08	0,02	0,00	0,09	0,01	0,00
14	0,60	0,40	0,00	0,15	0,08	0,00	0,43	0,18	0,00
15	0,60	0,20	0,00	0,45	0,06	0,00	0,45	0,09	0,00
16	0,90	0,40	0,00	0,65	0,35	0,00	0,43	0,19	0,00
17	0,20	0,00	0,00	0,03	0,01	0,00	0,24	0,02	0,00
18	0,10	0,10	0,00	0,01	0,01	0,00	1,00	1,00	0,00
19	0,10	0,00	0,00	0,01	0,00	0,00	0,20	0,01	0,00
20	0,90	0,20	0,00	0,34	0,24	0,00	0,24	0,10	0,00
21	0,80	0,50	0,00	0,54	0,35	0,00	0,53	0,27	0,00
22	0,90	0,50	0,00	0,68	0,39	0,00	0,46	0,18	0,00
23	0,80	0,10	0,00	0,46	0,19	0,00	0,37	0,13	0,00
24	1,00	0,80	0,00	0,24	0,09	0,00	0,91	0,32	0,00
25	0,40	0,20	0,00	0,18	0,03	0,00	0,28	0,04	0,00
26	1,00	1,00	0,00	0,47	0,33	0,00	0,89	0,60	0,00
27	0,80	0,60	0,00	0,29	0,17	0,00	0,64	0,33	0,00
28	0,40	0,20	0,00	0,10	0,06	0,00	0,31	0,12	0,00
29	0,50	0,20	0,00	0,28	0,21	0,00	0,28	0,17	0,00
30	0,90	0,40	0,00	0,73	0,25	0,00	0,62	0,11	0,00
31	1,00	0,30	0,00	0,39	0,20	0,00	0,24	0,08	0,00
32	1,00	0,50	0,00	0,85	0,52	0,00	0,41	0,10	0,00
33	0,50	0,10	0,00	0,27	0,17	0,00	0,26	0,11	0,00
34	0,20	0,00	0,00	0,17	0,04	0,00	0,16	0,03	0,00
35	1,00	1,00	0,00	0,98	0,68	0,00	0,73	0,29	0,00
36	1,00	1,00	0,00	0,99	0,91	0,03	0,80	0,70	0,00
37	1,00	0,80	0,00	1,00	0,72	0,00	0,94	0,63	0,00
38	1,00	0,60	0,00	0,90	0,24	0,00	0,47	0,11	0,00
39	0,90	0,40	0,00	0,64	0,25	0,00	0,34	0,06	0,00
40	0,90	0,40	0,00	0,58	0,13	0,00	0,27	0,03	0,00
41	0,90	0,80	0,00	0,89	0,68	0,00	0,52	0,34	0,00
42	1,00	0,80	0,00	0,77	0,58	0,00	0,29	0,14	0,00
43	1,00	0,80	0,00	0,45	0,25	0,00	0,16	0,12	0,00
44	1,00	0,70	0,00	0,80	0,53	0,00	0,37	0,12	0,00
45	1,00	0,10	0,00	0,67	0,07	0,00	0,38	0,01	0,00
46	1,00	0,90	0,00	0,80	0,62	0,00	0,50	0,35	0,00
47	0,70	0,50	0,00	0,39	0,32	0,00	0,14	0,08	0,00
48	1,00	0,40	0,00	0,69	0,21	0,00	0,48	0,13	0,00
49	1,00	0,50	0,00	0,26	0,10	0,00	0,30	0,12	0,00
50	0,70	0,30	0,00	0,50	0,22	0,00	0,13	0,04	0,00