FreePub: Collecting and Organizing Scientific Material Using Mindmaps

(or ...where mindmaps meet web search)

Theodore Dalamagas[1], Tryfon Farmakakis[2], and Manolis Maragkakis[3] and
Artemis G. Hatzigeorgiou[3]

[1] IMIS Institute/"Athena" R.C., Athens, Greece
`dalamag@imis.athena-innovation.gr`
[2] University of Edinburgh, School of Informatics, Edinburgh, Scotland
`T.Farmakakis@sms.ed.ac.uk`
[3] BSRC Fleming, Vari, GR
`(maragkakis,artemis)@fleming.gr`

Abstract. This paper presents a creativity support tool, called `FreePub`, to collect and organize scientific material using mindmaps. Mindmaps are visual, graph-based represenations of concepts, ideas, notes, tasks, etc. They generally take a hierarchical or tree branch format, with ideas branching into their subsections. `FreePub` supports creativity cycles. A user starts such a cycle by setting up her domain of interest using mindmaps. Then, she can browse mindmaps and launch search tasks to gather relevant publications from several data sources. `FreePub`, besides publications, identifies helpful supporting material (e.g., blog posts, presentations). All retrieved information from `FreePub` can be imported and organized in mindmaps. `FreePub` has been fully implemented on top of FreeMind, a popular open-source, mindmapping tool.

1 Introduction

Web search engines are widely used for searching information on the Web. Their increased popularity is due to the following reasons: the search model employed (i.e., keyword-based) is simple and easy to use, and the search techniques are nowadays mature enough to support fast text retrieval with accurate results.

However, there are use cases where the information need is complex. Consider, for instance, a researcher that needs to set up her research agenda and generate innovative ideas. She often has the "big picture" of the domain, i.e., an abstraction based on topics, thoughts, and everything else that helps setting up her search plan to explore the domain. Based on this initial abstraction, she (a) gathers information from several data sources, (b) organizes the information, (c) generates hypothesis and scientific results, (c) disseminates those results, and then (d) starts over by refining her abstraction and search plan. Such a *creativity cycle* actually enables discovery and innovation.

To illustrate an example of a creativity cycle, consider a researcher interested in *sequence matching techniques* for genomics, and the following use case:

1. The researcher starts by looking for journal papers that make a thorough review of this particular research area (i.e., the so-called survey papers), and blog articles that provide a review of the current state-of-the-art technologies technologies.
2. After organizing and studying the retrieved material, she pays more attention to the *local alignment* problem, that is "given a query sequence and a data sequence, find pairs of similar subsequences chosen from these sequences". She finds out that the dynamic programming solutions suggested to deal with that problem have high computational cost, and that this is the reason for researchers to work on approximation solutions (i.e., methods to return some but not all of the alignment results, according to some statistical significance model). Thus, she starts now looking for papers related to *approximate local alignment*.
3. After organizing and studying the retrieved material, she concludes that those methods, athough efficient, are not appropriate for several cases where the full result set of alignments is needed. Thus, she starts now looking for papers that are related to *indexing schemes for efficient local alinment*. These approaches exploit data structures which speed up the matching process between a large data sequences and a query sequence, at the expense of having to maintain these structures when data changes.
4. At any step of the above creativity cycle, she disseminate her findings to other researchers to get feedback.

New search models and techniques are necessary to support creativity and innovation [21]. A critical objective is to support creativity cycles, and also to provide effective presentation and visualization capabilities for the lists of retrieved resources that will guide users during their search and exploration.

Mindmapping [5, 10] makes use of visual diagrams to capture and organize information. They generally take a hierarchical or tree branch format, with ideas branching into their subsections. Mindmapping elements include concepts, ideas, notes, tasks, etc. One can use mindmaps to summarizing information, consolidating information from different research sources, thinking through complex problems, and presenting information showing the overall structure of her topic. Mindmaps is an excellent model for visualize, structure, and classify ideas, and support creative thinking.

This paper presents a creativity support tool, called `FreePub`, to collect and organize scientific material using mindmaps. `FreePub` supports creativity cycles, assisting users to:

1. set up their domain of interest using mindmaps,
2. browse mindmaps and launch search tasks to gather relevant documents from several data sources,
3. identify supporting material for those documents (e.g., blog posts, presentations), and
4. import and organise all retrieved information in mindmaps.

`FreePub` (http://web.imis.athena-innovation.gr/projects/mm/) has been built on top of Freemind [12], a popular open-source, mindmapping tool.

Outline. In the next section we give an overview of FreePub architecture, and we discuss the related work. Section 3 describes mindmaps. Section 4 presents the search facilities of FreePub, and Section 5 describes the semantic query expansion mechanism. Section 6 discusses a test case for FreePub, and, finally, Section 7 concludes the work.

2 Overview and Related Work

In this section we give a brief overview of tool features and technologies used, and we discuss the related work.

Figure 1 shows the architectute of FreePub. FreePub has been implemented on top of FreeMind [12]. Freemind provides an excellent user-friendly editor to build mindmaps. Users exploit mindmaps to set up their knwoledge domain, and collect and organize scientific material retrieved from several data sources. The *search orchestrator module* is responsible for launching *vertical* and *horizontal* search tasks, and coordinate their operation in order to retrieve publications and supporting material. The *semantic query expansion module* provides intelligent retrieval facilities by enriching user queries with terms extracted from mindmap elements to improve search effectiveness. The *data cleaning module* processes the result lists to remove name ambiguities and inconsistencies, and also to remove duplicate results. FreePub maintains a database of conference/journal info to assist cleaning tasks. The *facet-based browsing module* provides visualization options using several information facets to present the results. Finally, the *MM element construction module* is responsible for transfering the result lists into the mindmaps, according to user needs.

The use of mindmaps in information retrieval tasks has been acknowledged by several researchers. In [2], the authors present how information retrieval on mind maps could be used to enhance expert search, document summarization, keyword based search engines, document recommender systems and determining word relatedness.

Also, [3] describes how one can use mindmaps to succesfully model, design, modify, import and export XML DTDs, XML schemas and XML dooc, getting very manageable, easily comprehensible, folding diagrams. They actually converted a general purpose mind-mapping tool into a very powerful tool for XML vocabulary design and simplification. Finally, SciPlore MindMapping [1] is the first mind mapping tool focusing on researchers needs by integrating mind mapping with reference and pdf management. SciPlore MindMapping offers all the features one would expect from a standard mind mapping software, plus the following special features for researchers: adding reference keys, PDF bookmark import, and monitoring folders for new pdfs.

Compared to the above works, FreePub provides a full-fledged retrieval service to collect scientific material using mindmaps. It retrieves not only relevant publications, but also supporting material, like blog posts, presentation slides, from several wrapped data sources. Also, it exploits a semantic query expansion

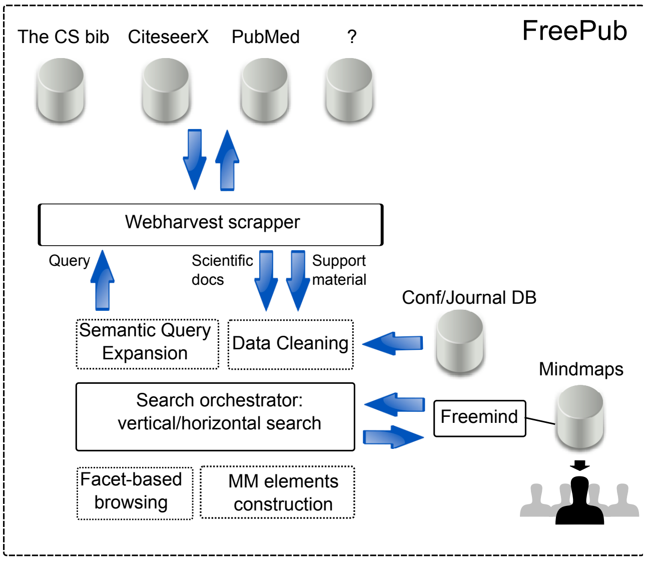


Fig. 1. FreePub's architecture.

mechanism to enrich user queries with mindmap element terms for improved search effectiveness.

There are also several open source (e.g., Vue, XMind, Compendium[4]) and commercial tools (e.g., MindManager, ConceptDraw, iMIndMap[5]) for mindmapping. However, they are actually mindmapping editors, providing advanced visualization capabilities, document handling and integration facilities with other popular software suites. Neither of them exploits mindmaps as a means for exploration Web search, giving also intelligent query expansion mechanisms, like FreePub does.

3 Mindmapping

Mindmapping [5, 10] refers to graphical representations of *elements* such as concepts, ideas, notes, tasks, or other items related to a topic of study. Mindmapping elements are organized in hierarchical branches or groups according to the semantic interpretation given by the user. However, everything is built around a central topic or idea. The key feature of mindmapping is that the elements are arranged in a non-linear fashion. Thus, users are free to enumerate and connect concepts without a tendency to begin within a particular conceptual framework.

[4] http://vue.tufts.edu/, http://www.xmind.net/, http://compendium.open.ac.uk/

[5] http://www.mindmanager.com, http://www.conceptdraw.com, http://www.thinkbuzan.com

This encourages a brainstorming approach to planning and organizational tasks, and idea generation.

Mindmaps is an excellent model for setting up workspaces for internet search, project and task management (including links to necessary files, executables, source of information), knowledge base organization (notes, references), and essay writing and brainstorming. They allow for greater creativity when recording ideas and information, and help the note-takers to associate topics and ideas with visual representations.

A key difference between mindmaps and other graph-based formal modelling representations, e.g. UML, semantic networks, TopicMaps, is that the the latter have explicit structured elements to model relationships. Contrary, mindmaps rerpesent the *visual mnemonics* of users, exploiting colors, icons and informal visual representations. Visual methods like mindmaps have been used for centuries in learning and problem solving by educators for recording knowledge, visual thinking, and problem solving. Also, mindmaps are based on radial hierarchies showing connections with a centered ruling concept.

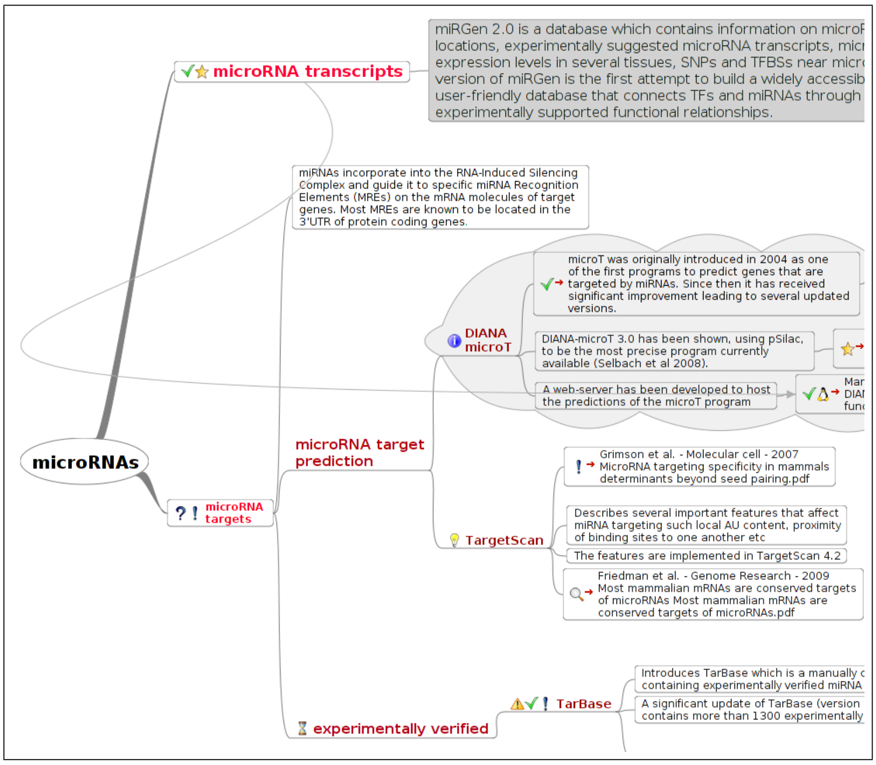
Freemind [12] provides a user-friendly editor to build mindmaps. Table 1 presents the most important mindmap elements used by Freemind. Figure 2 shows a mindmap example, organizing information about microRNA entities (see also Section 6). In this mindmap, for example, `microRNA` is the central idea where all other elements are structured around. `microRNA targets` and `microRNA transcripts` are topic elements, while `microRNA target prediction` is a subtopic element. The text "miRNA incorporate into the RNA-Induced..." is a detail element.

Topic, Larger topic: main elements, arranged in a topic/subtopic fashion, to represent ideas	**Waiting topic**: a topic that needs to be reconsidered
Needs action: an element for which action is needed	**Hot**: a critical element
Detail: text content element (e.g., notes, abstracts, etc)	**Link**: direct link element to user folders, urls, or local files
Object (keywords): set of words used as keywords for an element	**Object (code)**: piece of code for an element
Question: issues that need to be considered for an element	**Cloud**: set of related elements

Table 1. Mindmap elements in FreeMind.

4 Searching facilities

As the user explores a mindmap, she can initiate a search task to retrieve, from several wrapped data sources, documents relevant to mindmap topics. Various search parameters can be determined, like the number of results, the data



miRGen 2.0 is a database which contains information on microR
locations, experimentally suggested microRNA transcripts, micr
expression levels in several tissues, SNPs and TFBSs near micr
version of miRGen is the first attempt to build a widely accessib
user-friendly database that connects TFs and miRNAs through
experimentally supported functional relationships.

miRNAs incorporate into the RNA-Induced Silencing Complex and guide it to specific miRNA Recognition Elements (MREs) on the mRNA molecules of target genes. Most MREs are known to be located in the 3'UTR of protein coding genes.

microT was originally introduced in 2004 as one of the first programs to predict genes that are targeted by miRNAs. Since then it has received significant improvement leading to several updated versions.

DIANA-microT 3.0 has been shown, using pSilac, to be the most precise program currently available (Selbach et al 2008).

A web-server has been developed to host the predictions of the microT program

Mar
DIA
fun

Grimson et al. - Molecular cell - 2007
MicroRNA targeting specificity in mammals determinants beyond seed pairing.pdf

Describes several important features that affect miRNA targeting such local AU content, proximity of binding sites to one another etc

The features are implemented in TargetScan 4.2

Friedman et al. - Genome Research - 2009
Most mammalian mRNAs are conserved targets of microRNAs Most mammalian mRNAs are conserved targets of microRNAs.pdf

Introduces TarBase which is a manually c
containing experimentally verified miRNA

A significant update of TarBase (version
contains more than 1300 experimentally

Fig. 2. A mindmap example.

sources used, etc. For each search task, FreePub starts the retrieval service by first formulating the necessary queries. Keywords are extracted from the content of mindmap elements selected by the user in order to form keyword queries to send to the data sources. A key feature of FreePub is a semantic query expansion mechanism used to extract keywords not only from the selected mindmap elements, but also from their *semantic neighbourhood*. We discuss this feature in detail later on, in Section 5.

Vertical search. Keyword queries are sent to all wrapped data sources to retrieve relevant documents. Such data sources usually provide *vertical search facilities*, i.e., tailored to certain types of information resources - in our case, computer science publications (e.g., DBLP, PubMed [8, 19]). FreePub wraps data sources using WebHarvest [22]. We discuss wrapping facilities later on.

The resulting snippets are extracted from the data sources, cleaned, and presented to the user. Cleaning includes several facilities used to process the results in order to remove ambiguities, inconsistencies, etc. Specifically, the system utilizes a catalog with journal names and conferences extracted from DBLP and PubMed [8, 19] to deal with name inconsistencies. Each journal/conference name in the snippets is matched again this catalog to determine a common name for all snippets. The catalog actually maintains two string values for each journal/catalog entry: a short string for the acronym and a long one for the title of the entry.

Matching is based on the Levenshtein distance [14] L between two strings. The Levenshtein distance is defined as the minimum number of edit operations needed to transform one string into the other, with the allowable edit operations being insertion, deletion, or substitution of a single character. For example, L("VLDD", "VLDB Conf")= 6: replace 'D' with 'B', and insert ' ', 'C', 'o' 'n' 'f', a total number of 6 operations.

Assuming a string s and a catalog of n entries $\{(a_1, t_1), (a_2, t_2), \ldots, (a_n, t_n)\}$ with pairs of acronyms a_i and titles t_i, s is matched to the entry (a_i, t_i) such that $L(s, a_i) + L(s, t_i)$ is minimized $(0 < i \leq n)$. For example, "Very Large Database Conf" and "VLDB Conf", both are matched to ("Very Large Database Conference", "VLDB") catalog entry.

Duplicate elimination. Since results are retrieved from several data sources, duplicate results may appear. Duplicates are removed using entity resolution blocking techniques [23]. The problem of entity resolution involves finding records in a dataset that represent the same real-world entity. Blocking techniques divide data into groups and only compares records within the same group, to avoid redundant comparisons. This is based on the assumption that records in different blocks are unlikely to match.

FreePub implements the following efficient strategy for entity identification and duplicate elimination:

1. The result list of each data source is partitioned into groups, using the publication date as key for each group. For each group we maintain a `(key→value)` hash structure H, where `key` is the date and `value` is the list of publication objects o_i. For example: $H_1 = (2004 \rightarrow \{o_1, o_3, o_5, o_6\})$, $H_2 = (2005 \rightarrow \{o_2, o_4\})$ for data source 1, $H_3 = (2004 \rightarrow \{o_1, o_5, o_8\})$ for data source 2, etc.
2. Then, to identify duplicates we check pairs of publication objects (o_i, o_j) only for objects than share the same key (date). Checking is done using exact string matching on publication title and publication forum. For instance, in the previous example, only pairs of publication objects from H_1 value list and H_3 value list will be checked.

Horizontal search. After retrieving docuements relevant to mindmap elements, the user may launch another search task to get supporting material for these documents. Such material includes blog posts discussing the topic of a document, related presentations, other reports etc. To detect the material, `FreePub` uses *horizontal search facilities*, i.e., general search engines that cover all the Web, and appropriate options to restrict searches to only certain type of documents. Specifically, FreePub searches for the following support material for each retrieved publication:

1. pub document: a query string is constructed from publication's title, and the filetype:pdf or doc option is used in order to retrieve results. Further

heuristic rules are used in order to certify that the retrieved result is indeed the document of the publication. E.g., we parse the retrieved documents and check whether the title of the publication appears in, etc.

2. pub abstract: the abstract is extracted either by parsing the document identified in 1. or by looking for the appropriate metadata fields in the data source used, since several data sources provide such information.
3. slide presentation: a query string is constructed from publication's title, and the filetype:ppt or pdf option is used in order to retrieve results. Further heuristic rules are used in order to certify that the retrieved results are indeed presentations. E.g., we parse the retrieved documents and check whether certain terms appear inside, e.g., the term"outline", terms from the sections of the document identified in 1., etc.
4. blog entries: a query string is constructed from publication's title along with author's name and issued to the Google Blogs Search Engine to retrieve results.

Wrappers. `FreePub` retrieves scientific documents from several data sources, e.g., the collection of Computer Science Bibliography [7], citeseerX [6], and PubMed [19]. New data sources can be easily integrated. `FreePub` wraps data sources using WebHarvest [22], a Web scraping tool that (a) captures data source search capabilities, and (b) simplifies Web information extraction from data sources. WebHarvest provides several types of processors (e.g., html-to-xml, xpath, etc) to define a sequence of extraction operations on Web pages and identify the required html parts easily.

To demonstrate how WebHarvest work, we show the part of the html source of the first three results returned from google blog search for the term "ubuntu".

```
...1st result
<a href="http://www.howtoforge.com/how-to-upgrade-ubuntu-10.04-..." id="p-1">
How To Upgrade <b>Ubuntu</b> 10.04 (Lucid Lynx) To 10.10 (Maverick Meerkat)
(Desktop; Server)<br>
</font>
<font size=-1>
...2nd result
<a class=f1 href="http://www.howtoforge.com/" id="pb-1"
title="http://www.howtoforge.com/">
HowtoForge - Linux Howtos and Tutorials - -
http://www.howtoforge.com/</a>
</font>
</td>
</tr>
</table>
<p class=g></p>
...3rd result
<a href="http://www.readwriteweb.com/cloud/2010/10/latest-ubuntu-1010-emphasizes.php"
id="p-2">
Latest <b>Ubuntu</b> 10.10 Emphasizes the Cloud - ReadWriteCloud</a>
<table border=0 cellpadding=0 cellspacing=0><tr><td class=j>
<font color=#555555 size=-1>11 hours ago </font>
<font color=#555555 size=-1>by Audrey Watters</font><br><font size=-1>
Open source operating system <b>Ubuntu</b> 10.10 is available to download today for desktop,
notebook, and server editions. Hooray for well-timed 10.10;s. All these versions are
emphasizing Canonical embracing
...
```

WebHarvest is based on an XML configuration file describing the process to extract data. The elements define access to html pages, files, databases, mails, ftp servers and configures the work flow. An example of an XML configuration file that parses the above html source follows:

```
1.  <?xml version="1.0" encoding="UTF-8"?>
2.  <config charset="UTF-8">
3.  <var-def name="searchQuery" overwrite="false"/>
4.  <var-def name="content">
5.  <html-to-xml>
6.   <http url="http://blogsearch.google.com/blogsearch?hl=
       en&amp;oi=spell&amp;ie=UTF-8&amp;q=\${searchQuery}&amp;btnG=Search+Blogs"/>
7.  </html-to-xml>
8.  </var-def>
9. <var-def name="results1">
10.  <xpath expression="//a[contains(@id,’p-’)]">
11.   <var name="content"/>
12.  </xpath>
13. </var-def>
14. <var-def name="results2">
15.  <xpath expression="//td[@class=’j’]">
16.   <var name="content"/>
17.  </xpath>
18. </var-def>
19. </config>
```

In line 3, the variable `searchQuery` is assigned the value “ubuntu”, which is actually the search term. In line 6, the value of `searchQuery` is appended to the Google blogs search address and passed to the WebHarvests HTTP engine which returns the results page in raw HTML. In line 5, WebHarvests HTML-to-XML engine is called, which transforms the raw HTML code into a well formed XML document, which is assigned to the newly defined variable `content` in line 4. An abstract of the XML document that contains the information for one result is shown below:

```
...
<a href="http://www.readwriteweb.com/cloud/2010/10/latest-ubuntu-1010-emphasizes.php"
  id="p-1">Latest <b>Ubuntu</b>10.10 Emphasizes the Cloud - ReadWriteCloud</a>
  <table border="0" cellpadding="0" cellspacing="0">
    <tbody>
...
```

As we can see in the above excerpt, all the information we need for title and address is included in the first <a ... /a> line. To parse the information, in line 10, the WebHarvests XPath engine is called with the XPath expression //a[contains(@id,“p-”)] as argument which returns the title of the result. Similarly, in lines 14-18, we acquire the abstract of the result.

The advantage of using scraping tools to wrap Web data sources is that they simplify the interfacing with the data sources, since no hardcoded text processing code in needed. While technologies like Web services have become popular nowadays, scraping tools will always be necessary to get information form data sources that isnt yet offered through some SOAP-like interface.

Presentation and visualization. `FreePub` provides several facet-based visualization and presentation options to manipulate the resulting list of documents and their support material. The results may be organized by date, forum, author, or using any regular expressions that involves any of the above fields. Note that any time during a creativity cycle, the user may import any of the result (i.e., document, support material, etc) into the mindmap.

5 Semantic query expansion

In `FreePub`, query formulation is performed by extracting keywords from mindmap elements. The whole task is coordinated by a *s*emantic query expansion mechanism. The key point is that keywords are not extracted only from user-selected mindmap elements, but also from their *semantic neighbourhood*.

Initially, the semantic neighbourhood is decided automatically by the system, and includes important elements which are connected with the selected elements in the mindmap. The user may refine the neighbourhood by marking/unmarking mindmap elements.

`FreePub` employs a *term ranking scheme* to determine the top-K important terms (i.e., keywords) in the semantic neighbourhood of user-selected mindmap elements. These terms are used to expand the initial keyword query. Term importance is decided based on a tf/idf-oriented weighting scheme [11]. Terms are ordered accoring to their importance and the top-K terms are selected to expand the initial query. See for example Figure 3, where the user has selected the mindmap element "How to improve clustering" (marked by the system using a blue flag). Note that the system has also marked other mindmap elements around (marked using a green flag). These latter elements form the semantic neighbourhood of the selected element. Finally, the terms considered by the system for the query expansion are "clustering improve rank-based similarity". Next

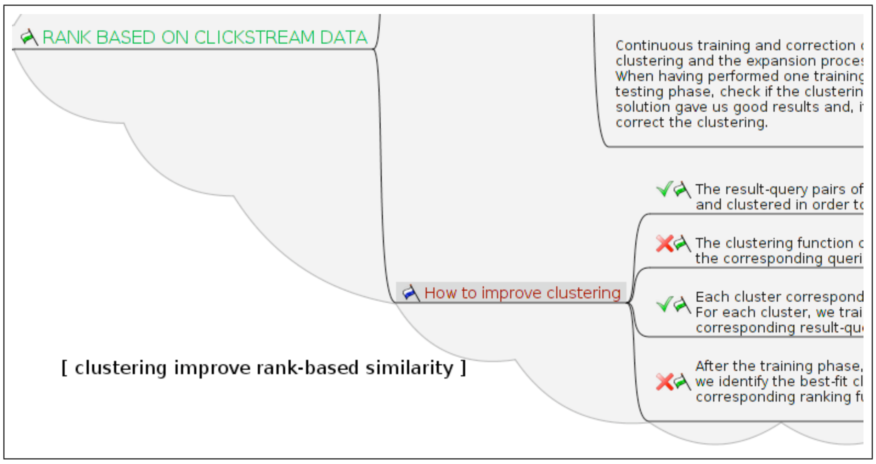

RANK BASED ON CLICKSTREAM DATA
Continuous training and correction
clustering and the expansion proce
When having performed one training
testing phase, check if the clusterin
solution gave us good results and,
correct the clustering.
✓ The result-query pairs of
and clustered in order t
How to improve clustering
✗ The clustering function
the corresponding queri
✓ Each cluster correspond
For each cluster, we trai
corresponding result-qu
✗ After the training phase,
we identify the best-fit cl
corresponding ranking f
[clustering improve rank-based similarity]

Fig. 3. An example of semantic query expansion.

we describe in detail how we determine the query expansion terms:

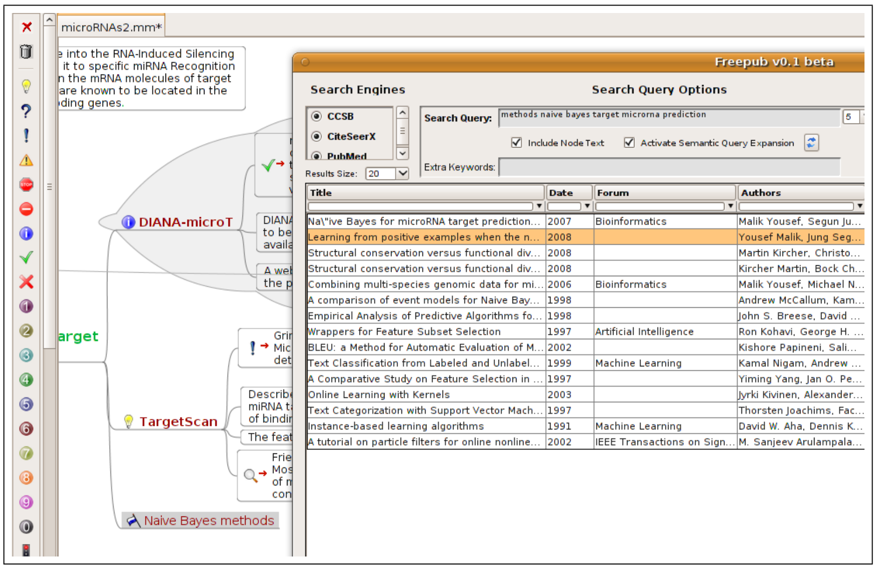
1. All elements in the neighbourhood of user-selected elements are considered as documents and are indexed using the Lucene IR engine [16]. The level of neighbourhoud is user defined, e.g., level 1 means that the neighbourhood of a selected element includes only directly adjacent nodes.
2. To each document d, we assign weights $docWeight_d$ according to the type of corresponding elements. For example, a document that is formed from topic elements gets higher weight than that formed from detail elements (see Table 1).
3. Terms are cleaned (i.e., punctuation and stopwords are removed), and the number of terms $docSize_d$ for each document d is calculated.
4. For each term t, we compute its number $freq_t^d$ of occurences in each doc d (i.e., term frequency), and the number $docFrec_t$ of documents containing term t.
5. Then, we compute, for each term t, its score w_t^d for every document d: $w_t^d = \frac{freq_t^d \times docFreq_t \times docWeight_d}{docSize_d}$. The final score W_t for each term t is the average of its scores w_t^d.
6. Terms are sorted according to W_t, and the terms with the better K scores are used to expand the initial query. K is user-defined.

6 FreePub in use

Since there are no mindmap benchmarks, we demonstrate FreePub advantages by presenting in this section a test case of working with FreePub (arranged with the research team of DIANA lab[6] at BSRC Fleming) to collect and organize scientific material regarding the microRNA target prediction problem.

Next, we give some background info for microRNAs to better understand the mindmap in Figure 2. microRNAs (miRNAs) are short RNA molecules that regulate gene expression by binding directly and preferably to the 3' untranslated region (3'UTR) of the sequence of genes [9]. Each mature miRNA is 19-24 nucleotides in length, and is processed from longer 70-nucleotide stem-loop structures known as pre-miRNAs. Pre-miRNAs are processed to mature miRNAs in the cytoplasm by interaction with the endonuclease Dicer. Each miRNA is integrated into the RISC (RNA induced silencing complex) complex and guides the whole complex to the mRNA sequence of a gene, thus inhibiting translation or inducing mRNA degradation [15]. Since their initial identification, miRNAs have been found to confer a novel layer of genetic regulation in a wide range of biological processes. MiRNAs were first identified in 1993 [20] via classical genetic techniques in C. elegans, but it was not until 2001 that they were found to be widespread and abundant in cells [18]. This finding served as the primary impetus for the development of the first computational miRNA target prediction programs. DIANA- microT [17] and TargetScan [4] were the first algorithms to predict miRNA target genes in humans, and led to the identification of an initial set of experimentally supported mammalian targets. Such targets are now

[6] http://microrna.gr/



Search Engines

● CCSB
● CiteSeerX
● PubMed

Results Size: 20

Search Query Options

Search Query: methods naive bayes target microrna prediction

☑ Include Node Text ☑ Activate Semantic Query Expansion

Extra Keywords:

Title	Date	Forum	Authors
Na\"ive Bayes for microRNA target prediction...	2007	Bioinformatics	Malik Yousef, Segun Ju...
Learning from positive examples when the n...	2008		Yousef Malik, Jung Seg...
Structural conservation versus functional div...	2008		Martin Kircher, Christo...
Structural conservation versus functional div...	2008		Kircher Martin, Bock Ch...
Combining multi-species genomic data for mi...	2006	Bioinformatics	Malik Yousef, Michael N...
A comparison of event models for Naive Bay...	1998		Andrew McCallum, Kam...
Empirical Analysis of Predictive Algorithms fo...	1998		John S. Breese, David ...
Wrappers for Feature Subset Selection	1997	Artificial Intelligence	Ron Kohavi, George H. ...
BLEU: a Method for Automatic Evaluation of M...	2002		Kishore Papineni, Sali...
Text Classification from Labeled and Unlabel...	1999	Machine Learning	Kamal Nigam, Andrew ...
A Comparative Study on Feature Selection in ...	1997		Yiming Yang, Jan O. Pe...
Online Learning with Kernels	2003		Jyrki Kivinen, Alexander...
Text Categorization with Support Vector Mach...	1997		Thorsten Joachims, Fac...
Instance-based learning algorithms	1991	Machine Learning	David W. Aha, Dennis K...
A tutorial on particle filters for online nonline...	2002	IEEE Transactions on Sign...	M. Sanjeev Arulampala...

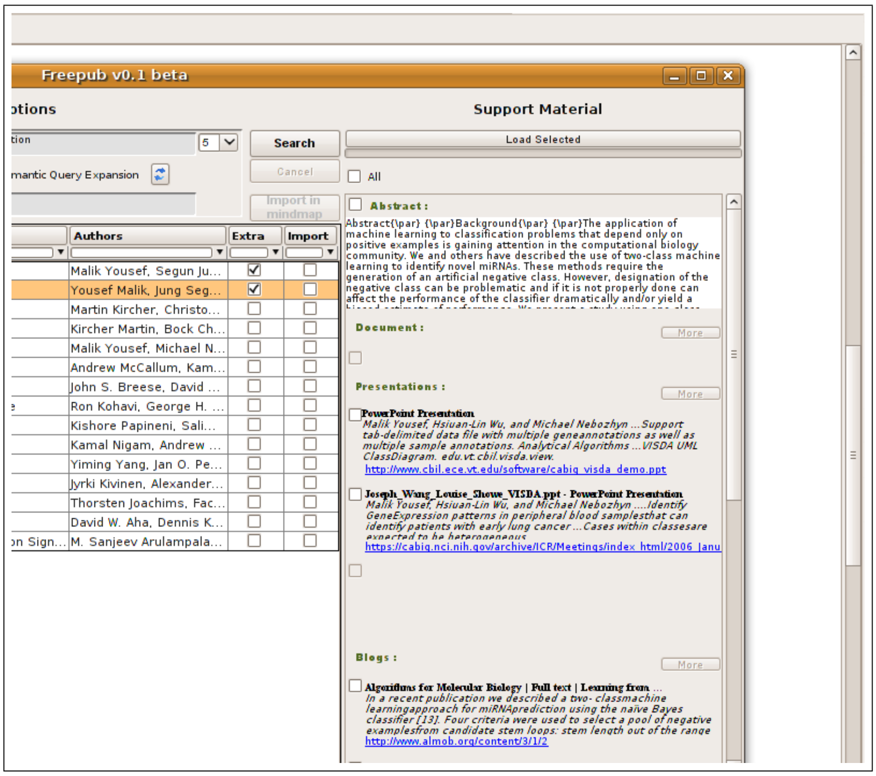
Fig. 4. Searching with `FreePub`.

collected and reported in TarBase [13] which contains more than one thousand entries for human and mouse miRNAs.

Figure 2 illustrates part of a mindmap for the miRNA target prediction problem set up by the researchers. Take for example the mindmap element **`microRNA target prediction`**, and its two subtopic elements `DIANA-microT` and `TargetScan`. Both predict genes that are targeted by miRNAs. The former was introduce in 2004, and since then it has received significant improvements. Currently has been shown (using pSilac) to be the most precise program currently available. The latter provides several important features that affect miRNA targeting.

Generally, most target prediction programs use several features to identify putative miRNA binding sites, such as evolutionary conservation, structural accessibility, nucleotide composition and others. Thus, a researcher considers that training learning functions using Naive Bayes models might be one way to follow for miRNA target prediction. She records this as a mindmap element, and starts the search. Figure 4 shows the resulting list of papers. Note that `FreePub` has expanded the initial user query from "Naive Bayes" to "methods naive bayes target microrna prediction", due to its semantic query expansion service.

The researcher selects, then, a couple of papers and a related presentation as supporting material to move to the mindmap. Figure 5 shows the retrieved supporting material, and Figure 6 shows the resulting mindmap.



	Authors	Extra	Import
	Malik Yousef, Segun Ju…	☑	☐
	Yousef Malik, Jung Seg…	☑	☐
	Martin Kircher, Christo…	☐	☐
	Kircher Martin, Bock Ch…	☐	☐
	Malik Yousef, Michael N…	☐	☐
	Andrew McCallum, Kam…	☐	☐
	John S. Breese, David …	☐	☐
e	Ron Kohavi, George H. …	☐	☐
	Kishore Papineni, Sali…	☐	☐
	Kamal Nigam, Andrew …	☐	☐
	Yiming Yang, Jan O. Pe…	☐	☐
	Jyrki Kivinen, Alexander…	☐	☐
	Thorsten Joachims, Fac…	☐	☐
	David W. Aha, Dennis K…	☐	☐
on Sign…	M. Sanjeev Arulampala…	☐	☐

☐ **Abstract :**

Abstract{\par} {\par}Background{\par} {\par}The application of machine learning to classification problems that depend only on positive examples is gaining attention in the computational biology community. We and others have described the use of two-class machine learning to identify novel miRNAs. These methods require the generation of an artificial negative class. However, designation of the negative class can be problematic and if it is not properly done can affect the performance of the classifier dramatically and/or yield a

Document :

☐

Presentations :

☐ **PowerPoint Presentation**
Malik Yousef, Hsiuan-Lin Wu, and Michael Nebozhyn ...Support tab-delimited data file with multiple geneannotations as well as multiple sample annotations. Analytical Algorithms ...VISDA UML ClassDiagram. edu.vt.cbil.visda.view.
http://www.cbil.ece.vt.edu/software/cabig_visda_demo.ppt

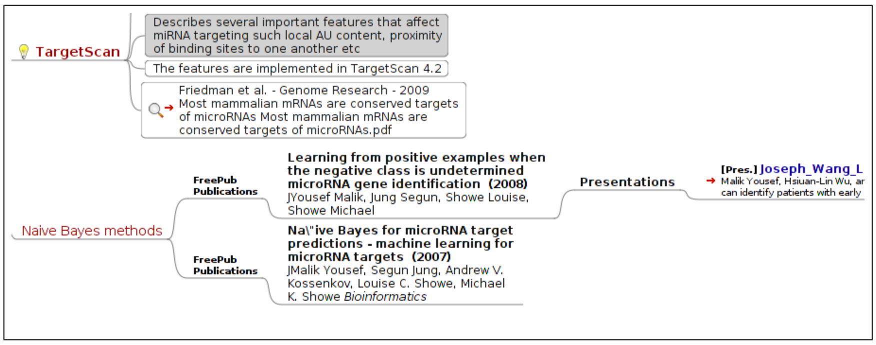
☐ **Joseph_Wang_Louise_Showe_VISDA.ppt - PowerPoint Presentation**
Malik Yousef, Hsiuan-Lin Wu, and Michael Nebozhyn ....Identify GeneExpression patterns in peripheral blood samplesthat can identify patients with early lung cancer ...Cases within classesare expected to be heterogeneous
https://cabig.nci.nih.gov/archive/ICR/Meetings/index_html/2006_Janu

☐

Blogs :

☐ **Algorithms for Molecular Biology | Full text | Learning from** ...
In a recent publication we described a two- classmachine learningapproach for miRNAprediction using the naïve Bayes classifier [13]. Four criteria were used to select a pool of negative examplesfrom candidate stem loops: stem length out of the range
http://www.almob.org/content/3/1/2

Fig. 5. Searching with FreePub: the supporting material.



Describes several important features that affect miRNA targeting such local AU content, proximity of binding sites to one another etc

The features are implemented in TargetScan 4.2

Friedman et al. - Genome Research - 2009 Most mammalian mRNAs are conserved targets of microRNAs Most mammalian mRNAs are conserved targets of microRNAs.pdf

Learning from positive examples when the negative class is undetermined microRNA gene identification (2008)
JYousef Malik, Jung Segun, Showe Louise, Showe Michael

[Pres.] Joseph_Wang_L
Malik Yousef, Hsiuan-Lin Wu, ar
can identify patients with early

Na\"ive Bayes for microRNA target predictions - machine learning for microRNA targets (2007)
JMalik Yousef, Segun Jung, Andrew V. Kossenkov, Louise C. Showe, Michael K. Showe *Bioinformatics*

Fig. 6. Importing results in mindmap.

7 Current status and future work

In this work, we presented FreePub, a creativity support tool to collect and organize scientific material using mindmaps. FreePub supports creativity cycles.

A user starts such a cycle by setting up her domain of interest using mindmaps. Then, she can browse mindmaps and launch search tasks to gather relevant publications from several data sources. `FreePub`, besides publications, identifies helpful supporting material (e.g., blog posts, presentations). All retrieved information from `FreePub` can be organized in mindmaps. `FreePub` has been fully implemented on top of FreeMind, a popular open-source, mindmapping tool.

For future work, we first plan to set up a detailed user-based evaluation of our tool with the help of a large number of scientists, and record their feedback after performing creativity cycles using `FreePub`. We also plan to develop several new services: (a) tagging facilities, (b) retrieval facilities for support material like, e.g., survey papers, highly-impact papers, etc., and (c) visual, easy-to-use scrapping facilities based on user query-by-example input in order to wrap data sources. Moreover, we will work on improving the semantic query expansion method. Finally, we will exploit public services like Mendeley and CiteULike[7] to evaluate the impact of retrieved publications, and the relations between them.

References

1. Jran Beel, Bela Gipp, and Christoph Mller. Sciplore mindmapping a tool for creating mind maps combined with pdf and reference management. *D-Lib Magazine*, 15(11), 2009.
2. Jran Beel, Bela Gipp, and Jan Olaf Stiller. Information retrieval on mind maps what could it be good for? In *Proceedings of the CollaborateCom09 Conference*, 2009.
3. Alejandro Bia, Rafael Muoz, and Jaime Gmez. Using mind maps to model semistructured documents, metadata, and semantic structures. In *Proceedings of the ECDL Conference (poster)*, 2010.
4. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, and Burge CB. Prediction of mammalian microrna targets. *Cell*, 115(7):787–798, 2005.
5. Tony Buzan. *The Mind Map Book.* Penguin Books, 1996.
6. CiteseerX. http://citeseerx.ist.psu.edu/.
7. The collection of Computer Science Bibliography. http://liinwww.ira.uka.de/bibliography/.
8. DBLP. http://www.informatik.uni-trier.de/˜ley/db/.
9. Bartel DP. Micrornas: genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, 2004.
10. Paul Farrand, Fearzana Hussain, and Enid Hennessy. The efficacy of the mind map study technique. *Medical Education*, 36(5):426–431, 2002.
11. Bill Frakes and Ricardo Baeza-Yates, editors. *Information Retrieval Data Structures & Algorithms.* Prentice Hall, 1992.
12. Freemind. http://freemind.sourceforge.net.
13. Papadopoulos GL, Reczko M, Simossis VA, Sethupathy P, and Hatzigeorgiou AG. The database of experimentally supported targets: a functional update of tarbase. *Nucleic Acids Research*, 37(1):155–158, 2008.
14. Dan Gusfield. *Algorithms on strings, trees, and sequences.* Cambridge University Press, 1999.

[7] http://www.mendeley.com/, http://www.citeulike.org/

15. Liu J, Carmell MA, Rivas FV, Marsden CG, Thomson JM, Song JJ, Hammond SM, Joshua-Tor L, and Hannon GJ. Argonaute2 is the catalytic engine of mammalian rnai. *Science*, 305(5689):1437–1441, 2004.
16. Lucene. http://lucene.apache.org.
17. Kiriakidou M, Nelson PT, Kouranov A, Fitziev P, Bouyioukos C, Mourelatos Z, and Hatzigeorgiou A. A combined computational-experimental approach predicts human microrna targets. *Genes Dev*, 18(10):1165–1178, 2004.
18. Lagos-Quintana M, Rauhut R, Lendeckel W, and Tuschl T. Identification of novel genes coding for small expressed rnas. *Science*, 294(5543):853–858, 2001.
19. PubMed. http://www.ncbi.nlm.nih.gov/pubmed/.
20. Lee RC, Feinbaum RL, and Ambros V. The c. elegans heterochronic gene lin-4 encodes small rnas with antisense complementarity to lin-14. *Cell*, 75(5):843–854, 1993.
21. Ben Shneiderman. Creativity support tools: accelerating discovery and innovation. *Communications of the ACM*, 50(12):20–32, 2007.
22. WebHarvest. http://web-harvest.sourceforge.net/.
23. Steven Euijong Whang, David Menestrina, Georgia Koutrika, Martin Theobald, and Hector Garcia-Molina. Entity resolution with iterative blocking. In *Proceedings of the ACM SIGMOD Conference*, 2009.