BioSumm: a novel summarizer oriented to biological information

Elena Baralis, Alessandro Fiori, Lorenzo Montrucchio

Abstract— **The availability of increasingly wider repositories of biomedical and biological texts requires effective techniques to manage the huge mass of unstructured information there contained. The availability of ad-hoc document summaries, targeted to specific topics, may assist researchers in inferring previously undisclosed knowledge and in performing the biological validation of the results of data mining analysis.**

This paper presents BioSumm, a flexible framework which analyzes large collections of unclassified biomedical texts and produces ad-hoc summaries oriented to inferring knowledge of gene/protein relationships. Summary generation is driven by a novel grading function, which biases sentence selection by means of an appropriate domain dictionary.

I. Introduction

In recent years, the growing availability of large document collections has stressed the need of effective and efficient techniques to operate on them (e.g., navigate, analyze, infer knowledge and represent it in the most suitable way). Given the huge amount of available information, it has become increasingly important to provide improved mechanisms to detect and present the most relevant parts of textual documents effectively. This becomes even more crucial in the life science domain in which huge quantities of data are steadily produced by researchers all over the world.

Initially, the task of analyzing the most relevant parts of texts and of performing on demand data integration for inferring new knowledge and for validation purposes was manually performed by molecular biologists [13]. This approach has become unfeasible, due to the huge amount of information that is daily generated and contributed by a vast research community spread all over the world. In fact, repositories like PubMed Central [7], the U.S. National Institutes of Health (NIH) free digital archive of biomedical and life sciences journal literature, nowadays contain billions of documents and are constantly growing.

Recently, many research efforts have been devoted to automatically indexing and managing the highly unstructured information contained in texts. Conventional "general purpose" information retrieval systems, including modern search engines, find and rank documents based on maximizing relevance to the user query [15]. However, they still require users to follow the hyperlinks, to read the documents and to locate the sentences that are more relevant for their information seeking goals. The iHop project [14] is explicitly tailored for biomedical articles. It uses genes and proteins as hyperlinks between sentences and abstracts and it converts the information in PubMed Central into one navigable resource. However, it provides only a link to the texts and leaves to the user the task of finding the most appropriate documents by browsing them.

Other works exploit text summarization [16], also tailoring it for the biomedical domain [19] and trying to better refine the produced summary by exploiting semantic information and ontology knowledge [23]. They provide to the user a more concise and compact version of the document. Thus, they better fulfill the need of reducing and organizing the huge amount of unstructured information contained in the texts. The sentences extracted by all these summarizers are suitable to provide a human readable synthesis and to emphasize the main ideas of an article or of a group of articles. However, these summarizers give only a general description of the major topics in the texts and tend to discard the most domain-specific sentences (e.g., the ones listing genes and their interactions). These sentences may instead be very important for biological validation and knowledge inference.

Other approaches [28] tackle the problem of sentence representation by means of graphs. However, they suffer from the same limitation of the previous summarization techniques. Furthermore, they are more suitable for collections of classified texts, that are only a subset of the available biomedical literature.

In this paper we present the BioSumm (Biological Summarizer) framework that analyzes large collections of unclassified biomedical texts and exploits clustering and summarization techniques to obtain a concise synthesis, explicitly addressed to emphasize the text parts that are more relevant for the disclosure of genes (and/or proteins) interactions. The framework is designed to be flexible, modular and oriented to biological information. Researchers can exploit BioSumm for knowledge inference and biological validation of the interactions discovered in independent ways (e.g., by means of data mining techniques).

The paper is organized as follows. Section II presents the architecture of the proposed framework and describes its main blocks. Section III discusses preliminary experimental results, while Section IV draws conclusions and presents future developments of this work.

II. Framework Description

The BioSumm framework processes biomedical texts for which no class labels and no division by topics are provided. It produces a good quality summary targeted to a specific goal, which in our case is inferring knowledge

Manuscript received July 5, 2008.

A. Fiori is with the Politecnico di Torino, Italy (corresponding author, phone: 0039 011 090 7194; fax: 0039 011 090 7099; alessandro.fiori@ polito.it)

E. Baralis and L. Montrucchio are with the Politecnico di Torino, email: elena.baralis@polito.it, lorenzo.montrucchio@studenti.polito.it

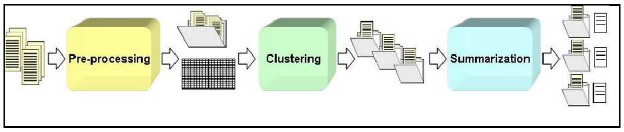

Fig. 1. Framework architecture

of gene/protein relationships. The summary is generated selecting from groups of similar texts the subset of sentences which better describe the group in the context of the selected goal.

The framework is characterized by a flexible and modular structure, shown in Figure 1 and composed by the following blocks:

- **Preprocessing.** It extracts relevant parts of the original document and performs text stemming.
- **Clustering.** It divides rather diverse texts into homogeneous clusters, in which the documents cover the same topic.
- **Summarization.** It produces a summary for each cluster.

The first two blocks are general purpose blocks and their goal is to prepare the document collection for the biological information extraction. The third block is specifically tailored to biological information. In the following subsections each block of the framework will be covered in details.

A. Preprocessing

The preprocessing block extracts the relevant information from the considered document sources. Many different biological document sources are available [1], [3], [6], [7]. For example, PubMed Central is a well known public repository for research articles. Such articles, all belonging to scientific journals, are downloadable free of charge [7] in the form of a .nxml file, which is XML for the full text of the article, encoded in the NLM Journal Archiving and Interchange DTD [4].

To build a common representation of texts, BioSumm performs two preprocessing steps: (i) extraction of relevant parts of the article from XML files, (ii) construction of the document matrix for the whole collection.

The original format provided by PubMed Central and designed for XML-based mining analysis, contains several tags (e.g., “journal” or “date of publication”) that are not meaningful for biological information retrieval. The first preprocessing step extracts from the XML files the relevant parts of research papers, namely title, abstract, body and, when available, the keywords that describe the content of the article. The user may select which parts should be used for the analysis. This step, given in input either semi-structured XML files or plain unstructured text files, produces a uniform (text) output.

The second preprocessing step produces a matricial representation W of a source in which each row is a document and each column corresponds to a feature (word) of the documents. Each element of matrix W is the TFIDF (term frequency - inverse document frequency) value for a term, computed as follows:

$$W_{ij} = \mathit{tf}_{ij} \cdot \mathit{idf}_j \tag{1}$$

where tf_{ij} is the term frequency of word j in document i and idf_j is the inverse document frequency of term j. The tf_{ij} term in (1) is defined as:

$$\mathit{tf}_{ij} = \frac{n_{ij}}{\sum_k n_{kj}} \tag{2}$$

where n_{ij} is the number of occurrences of the considered term in document i and the denominator is the number of occurrences of all terms in document i. Moreover, the idf_j term is defined as:

$$\mathit{idf}_j = \log \frac{|D|}{|\{d : j \in d\}|} \tag{3}$$

where $|D|$ is the number of documents in the collection and $|\{d : j \in d\}|$ is the number of documents in which term j appears.

Matrix W is generated by means of the text plug-in of RapidMiner [17]. First of all, it divides the text in chunks by means of a tokenizer. Then it filters all the produced chunks with an English stopword filter and with a token length filter that prunes the words shorter than two characters. Finally stemming is performed by exploiting the Porter stemming algorithm [22]. In most cases the generated matrix is still characterized by a high dimensionality. Hence, a further filtering part eliminates “useless features”, i.e., very frequent words that tend to be non discriminative in the clustering phase.

B. Clustering

This block divides unclassified texts, belonging to specialized journals, into more homogeneous subsets. The clustering phase is very important to detect texts which share a common topic without any a priory knowledge of their content. Without this step the quality of the summary decreases because there is no strong correlation between document topics. The clustering block performs its analysis on matrix W produced by the preprocessing block.

Clustering is performed by means of the CLUTO software package [2]. CLUTO clusters high-dimensional data and can scale to large datasets containing hundreds of thousands of objects and tens of thousands of dimensions. This is exactly the kind of scenario in which our framework operates more frequently. Furthermore, it produces a detailed list of the most distinctive features (words) of each cluster.

Since the document collections addressed in this work all belong to a common scientific context and share the same vocabulary, they are not strongly heterogeneous in terms of topics. Hence, CLUTO is configured to minimize computational time, because the quality of generated clusters is already appropriate for our needs.

Clustering is performed by an optimization process which seeks to maximize or minimize a particular clustering criterion function defined either globally or locally over the entire clustering solution space. Therefore, the main configuration choices are the clustering algorithm, the similarity measure, and the criterion function. BioSumm is based on a partitional algorithm, the repeated-bisecting method, which produces a globally optimized solution. This method reaches a suitable trade off between the quality of the results and the scalability guaranteed by partitional algorithms [21], [27]. The selected similarity measure is the cosine similarity function, which further improves the scalability of the approach. The combination of cosine correlation and repeated bisecting method is the most scalable in terms of time and space complexity [26], because its time complexity is $O(NNZ * log(k))$ and its space complexity is $O(NNZ)$, where NNZ is the number of non-zero values in the input matrix and k is the number of clusters. The selected criterion function is:

$$max \sum_{i=1}^{k} \sqrt{\sum_{v,u \in S_i} Sim(v,u)} \tag{4}$$

where k is the total number of clusters, S_i is the set of objects assigned to cluster i, v and u represent two objects and $Sim(v,u)$ is the similarity between the two objects. This criterion function is suitable in cases of high dimensionality and demanding scalability issues [26].

Since a partitional algorithm is used, the number of clusters is required as input. BioSumm allows the user to select it. The effect of different value selection is explored in Section III-B.

C. Summarization

This block is the core of our framework. It provides, separately for each cluster determined by the previous block, an ad-hoc summary, containing the sentences that are potentially more useful for inferring knowledge of gene/protein relationships.

Our ad-hoc summarizer is based on OTS (Open Text Summarizer) [5] a single-document summarizer whose implementation was proved to be particularly efficient by recent studies [25]. The Open Text Summarizer is an open source summarizing tool that ships with major Linux distributions. As many single-document summarizers, it is based on the idea that the most relevant sentences are those containing the largest number of the most frequent words in the document (stopwords excluded). These words are usually the ones that better describe the topics of the documents.

The BioSumm summarizer exploits the efficient structure of the original OTS. It scans the text once and stores in a sorted list terms (properly stemmed) and their frequencies. Then, the text is split into sentences and each sentence is graded. The sentences with the highest score are selected to build a summary, containing a given percentage of the original text. This percentage, which is set by the user, is called summarization ratio.

The core of the BioSumm summarizer is a novel grading function that takes into account the occurrences (i.e., the number of times a word appears in the document) of some domain specific words. These words are stored in a dictionary. In this work, to focus on gene/protein information, the dictionary contains human gene and protein names and aliases. The dictionary is built by querying the Biogrid publicly available database [20].

Let T be the set of all the terms in the texts, K a subset of T ($K \subseteq T$) which contains only the words that are not filtered by the stopword analysis and G the set of all genes and proteins in our dictionary. The grading function for sentence j in document i is given by

$$gf_{i,j} = \delta_j \cdot \sum_n tf_{in} \tag{5}$$

where tf_{in} represents the frequency of term n belonging to set K in document i and δ_j is a weighting factor which considers the number of occurrences of dictionary entries in sentence j. δ_j is defined by

$$\delta_j = \begin{cases} 1 & \text{if } tos_n = 0 \quad \forall n \in G \\ \alpha + \beta \cdot \sum_n tos_{g_n} & \text{otherwise} \end{cases} \tag{6}$$

where tos_n represents the number of occurrences, in sentence j, of term n belonging set G, and α and β are two constant factors. α belongs to the range $[1, +\infty)$ and its role is to favour the sentences that contain terms in G, disregarding their number. β is instead in the range $[0, 1]$ and weights the occurrences of words of G. With $\alpha = 1$ and $\beta = 0$ the summarizer ignores terms in G, thus disregarding the dictionary. By increasing α, the presence of a gene or protein of G raises the score of the sentence, but sentences with a different number of gene references are weighted identically. To weight the occurrences of terms of G, β should be different from 0. The closer β is to 1, the more different gene occurrences in the sentence are deemed relevant. After several experiments, we selected $\alpha = 2$ and $\beta = 0.1$. This setting selects the sentences referencing one or more genes, and also gives relevance to the number of gene occurrences to capture the gene/protein interactions.

The original OTS version exploits a simpler grading function which involves a constant multiplicative factor based on the "structure" of the document (e.g., the leading sentence of a new paragraph). This grading function is effective in producing a summary which is easily readable by humans. The summary covers the major topics of the document, but it does not necessarily contain sentences which would be relevant for our targeted search. For example, in a cluster of documents related to a given disease, the OTS summary contains general descriptions of the disease itself, but it tends to ignore the biological information (e.g., protein/protein interactions). Our grading function, while still selecting sentences more related to the major topics of the cluster, also favors the ones referencing the entries of the dictionary. Thus, it also includes sentences potentially more meaningful for further biological analysis.

III. Preliminary Experimental Results

We performed experiments on a subset of the PubMed Central [7] text collection. The relevant characteristics of this subset are described in Table I. When for all the articles in

TABLE I

DOCUMENT COLLECTION CHARACTERISTICS

Collection	Journal	Size	Keywords
Bioinformation	Bioinformation	160	NO
Breast_Cancer	Breast Cancer	467	YES
J_Key	Breast Cancer Arthritis Res	927	YES
Crit_Care	Crit Care	1460	NO

the collection the keyword field is available, the *"Keywords"* column in Table I is *"Yes"*, otherwise it is *"No"*. The collections are characterized by different cardinalities.

We performed three sets of experiments to evaluate:

- the capability of the summarizer to identify relevant information
- the capability of the clustering block to group similar documents
- the scalability of the various parts of the BioSumm framework

A. Summarization analysis

The purpose of our framework is to build a summary which captures the main biological features of the articles. This set of experiments tests the quality of the produced summaries. We focused the analysis on the Breast_Cancer collection, but similar results were obtained on the other collections.

We set the number of clusters to 80, to reach a reasonable trade off between computational time and quality of the result. This issue is further analyzed in Section III-B. Both abstract and body have been considered and the summarization ratio is set to 20% to obtain compact summaries.

We focus the analysis on one of the obtained clusters, which is composed by ten documents. The "keywords" of this cluster, namely the most descriptive and discriminative words for the cluster, are *proband*, *Ashkenazi*, and *Jewish*. A proband is the family member through whom a family's medical history comes to light, whereas Ashkenazi Jews, also known as Ashkenazic Jews or Ashkenazim are the Jews descended from the medieval Jewish communities of the Rhineland. Hence, the cluster likely deals with genetic peculiarities or diseases that occur in certain ethnic populations [24].

In Table II we report the six sentences graded with the highest scores by BioSumm and the six top sentences selected by OTS. BioSumm generally gives a high score to the sentences containing genes, which are very likely selected for the summary. More specifically, all top sentences contain at least a reference to BRCA1 or BRCA2, that are human genes belonging to a class known as tumor suppressors [10], [12]. Furthermore, among the sentences that contain these genes, the summarizer prefers those referencing the highest number of them. These sentences are more relevant for knowledge inference, because they may describe the relationship between several genes/proteins. For example, by considering the second sentence, we may learn that BRCA1 and BRCA2 are also involved in breast/ovarian cancer. We have the biological evidence of the correctness of this information in [9] and [11], which are scientific papers not belonging to our collections.

The sentences selected by BioSumm are all closely related to the keywords of the cluster (e.g., most sentences describe gene interactions discovered in statistical analysis on different populations). Hence, the BioSumm grading function, although strongly gene and protein oriented, is still capable of detecting the most relevant sentences for the topics of the cluster, which deals with genetic studies on populations. This is confirmed by considering the second column of Table II, which contains the sentences selected by OTS. The top two sentences are the same for both summarizers, while the fourth sentence extracted by OTS is exactly the third extracted by BioSumm. The third and the fifth sentences selected by OTS are long sentences that introduce new paragraphs and deal with statistical analysis, but not directly with biological topics. For this reason BioSumm discards them, while OTS selects them because of their position in the text. Finally, the sixth sentence is particularly important because it is a very short and technical sentence that tends to be pruned by most summarizers. BioSumm selects a sentence which is really meaningful for our purposes, because it describes a gene (BRCA1) and five of its mutations (described also in [8]). The sentence extracted by OTS, instead, albeit addressing the same issue, is more general and misses all the gene mutations, whereas our framework was able to capture this crucial piece of knowledge.

B. Clustering evaluation

The role of the clustering block is to divide a collection in small subsets, maximizing the internal similarity and cohesion of each cluster, without any a-priori knowledge of the document contents. Therefore, a good cluster is a group of documents sharing similar topics.

To measure the agreement between topics and clustering results we computed the Rand Index [18]. It measures the number of pairwise agreements between a clustering K and a set of class labels C over the same set of objects. It is computed as follows

$$R\left(C,K\right) = \frac{a+d}{a+b+c+d} \tag{7}$$

where a denotes the number of object pairs with the same label in C and assigned to the same cluster in K, b denotes the number of pairs with the same label, but in different clusters, c denotes the number of pairs in the same cluster, but with different class labels and d denotes the number of pairs with a different label in C that were assigned to a different cluster in K. The values of the index are in the range 0 (totally distinct clusters) and 1 (exactly coincident clusters). The Rand Index is meaningful for a number of clusters in the range $[2; N-1]$, where N is the number of objects. Moreover, clusters with only one element are penalized giving no contribution to Rand Index analysis.

We analyzed the J_Key collection, in which some keywords are available for all articles. The keywords provide an objective way to define the topics of the articles. We

TABLE II
SENTENCES WITH THE HIGHEST SCORE IN CLUSTER ”PROBAND, ASHKENAZI, JEWISH”

Rank	BioSumm sentences	OTS sentences
1)	In contrast to studies on North and East European populations the present results indicate a lack of relevant founder effects for BRCA1 and BRCA2 -related disease in the sample of patients analyzed, which is in agreement with other Italian studies and with ethnical and historical data.	In contrast to studies on North and East European populations the present results indicate a lack of relevant founder effects for BRCA1 and BRCA2 -related disease in the sample of patients analyzed, which is in agreement with other Italian studies and with ethnical and historical data.
2)	This is a low proportion compared with studies that suggested that BRCA1 and BRCA2 are responsible for the large majority of breast/ovarian cancer families, with the greater proportion due to BRCA1.	This is a low proportion compared with studies that suggested that BRCA1 and BRCA2 are responsible for the large majority of breast/ovarian cancer families, with the greater proportion due to BRCA1.
3)	Furthermore, BRCA2 and, to a lesser extent, BRCA1 also appear to be responsible for an important, but still debated proportion of male breast cancers.	Third, we let i Y = log(2 i p) if the i i th woman was a carrier and log[2(1-p)] otherwise, i E1 = n log2 + p log(ip) + (1-ip) log(1-ip) and i O1 = Y.
4)	Knowledge of the contribution of BRCA1 and BRCA2 to breast cancer in these patients is still incomplete.	Furthermore, BRCA2 and, to a lesser extent, BRCA1 also appear to be responsible for an important, but still debated proportion of male breast cancers.
5)	The overall proportion of cancer-affected males with BRCA2 mutations (10%) was high compared with data from other outbred populations, but was lower than that reported for populations with founder effects.	The statistic i Z1 = (O1-E1)/[var(E1)] 1/2 , where var(E1) = p(1-ip)log[ip/(1-ip)] i 2 has a standard normal distribution under the null hypothesis, and deviations test whether the predicted values were too clustered or too dispersed.
6)	The five deleterious BRCA1 mutations (Table 2) included four frameshift mutations (BRCA1 1479delAG, BRCA1 1623del5bp, BRCA1 3880delAG, BRCA1 5083del19bp) and one missense mutation (BRCA1 300TtoG).	These mutations were already reported in the literature or in the Breast Cancer Information Core electronic database.

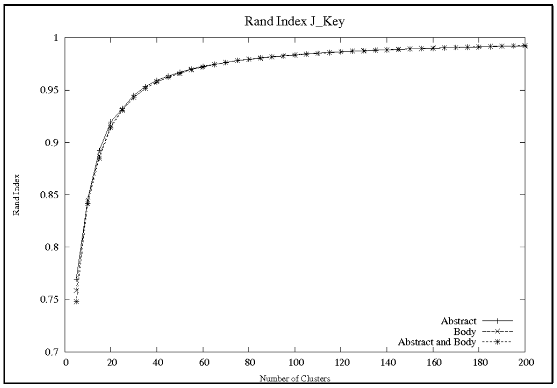

Fig. 2. Rand Index on J_key

clustered the keyword descriptors of the articles and we used the resulting clusters as class labels C for the Rand Index. Separately, we clustered the abstracts, the bodies, and the abstracts+bodies of the same documents. We repeated the experiment with several values of the cluster number parameter.

Figure 2 reports the results of the experiments. The Rand Index is generally high and becomes very close to 1 for more than 40 clusters, because smaller clusters (containing around 10-20 documents) tend to include more homogeneous documents. The clustering result of the keywords and the results obtained using the other parts of documents are very similar. Hence, the clustering block clusters the documents according to the topics they actually deal with. Similar findings were obtained with Breast_Cancer, the other collection provided with keywords.

C. Performance analysis

To evaluate the performance of BioSumm, we analyzed the completion times of the various framework blocks and their impact on the total completion time. All the four article collections, characterized by a different cardinality, have been considered. The framework scalability with the document number has also been analyzed. The analysis has been performed by considering both the abstract and the body of the documents. For each collection the most suitable values for the cluster number and summarization ratio parameters are also reported. Experiments were performed on an Intel Centrino Duo processor T2300 @ 1.66GHz with 2GByte of RAM. All reported execution times are real times, including both system and user time, and obtained from the unix time command. The performance for the four document collections is reported in Table III. The total time takes into account also the input/output among the blocks.

Preprocessing performance. The results show that the time required by the preprocessing block scales well with the number of documents. The reported performance also depends on the density of the collection and the size of the documents. Furthermore, the computational time of this step is roughly 20%-25% of the total time and the impact of the block decreases as the number of documents grows. We performed the same experiment with different settings of the number of clusters and summarization ratio parameters and we obtained similar results.

Clustering performance. In this analysis we set the summarization ratio to 20% and increased the number of clusters with constant increments. The results, reported in Figure 3, show that the clustering time scales well both with the number of documents in the collection and the number of clusters.

Summarization performance. The last set of experiments is focused on the summarization block. We analyzed the impact of the summarization ratio on performance. In this analysis we set the same number of clusters for all the collections. The analysis shows that the summarization ratio has no impact on performance, because the computational

TABLE III

PERFORMANCE OF THE BIOSUMM SUMMARIZER

Collection	Size	Cluster Number	Summarization Ratio (%)	Preprocessing time (sec)	Clustering time (sec)	Summarization time (sec)	Total time (sec)
Bioinformation	160	40	20	27.63	10.72	43.57	82.82
Breast_Cancer	467	80	20	109.85	80.72	235.82	430.77
J_key	927	100	20	259.40	288.47	512.32	1080.79
Crit_Care	1460	120	20	355.34	505.13	642.58	1544.30

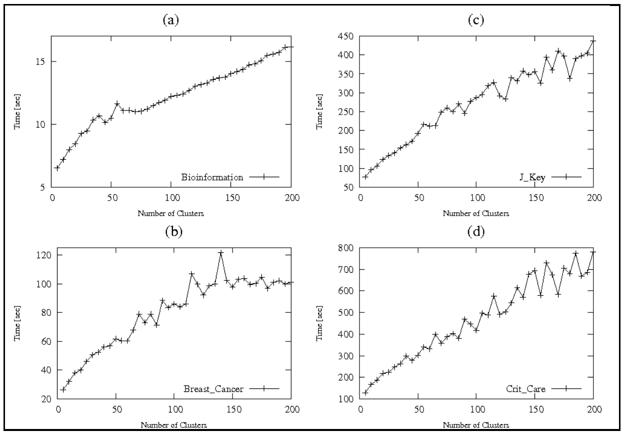

Fig. 3. Performances of the clustering block on (a) Bioinformation, (b) Breast_Cancer, (c) J_Key, (d) Crit_Care collection

times do not vary significantly when varying the ratio.

IV. CONCLUSIONS

BioSumm is a flexible and modular framework to generate ad-hoc document summaries oriented to biological content, in particular to gene and protein information. Preliminary experimental results show that BioSumm can summarize large collections of unclassified data by extracting the sentences that are more relevant for knowledge inference and biological validation of gene/protein relationships. Although focused on a specific subject, its capability to detect the sentences that better cover the major topics of a group of documents is still preserved. Researchers that discover gene correlations by means of analysis tools (e.g., data mining tools) may exploit this framework to effectively support the biological validation of their results.

As future works, we will evaluate the possibility of extending the summarization approach to multi-document summarizers. Furthermore, integration of ontology derived knowledge in the clustering phase will be considered. Finally, we will validate the effectiveness of our approach in different domains (e.g., financial articles).

REFERENCES

[1] Bioline international. http://www.bioline.org.br/.

[2] Cluto - software for clustering high-dimensional datasets. http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview.

[3] Ispub - internet scientific publications. http://www.ispub.com.

[4] Nlm journal archiving and interchange tag suite. http://dtd.nlm.nih.gov/.

[5] Open text summarizer. http://libots.sourceforge.net/.

[6] Plos - public library of science. http://www.plos.org/.

[7] Pubmed central ftp service. http://www.pubmedcentral.nih.gov/about/ftp.html#Source_files.

[8] A. Antoniou, P. Pharoah, and S. Narod. Average risks of breast and ovarian cancer associated with brca1 or brca2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. *Am J Hum Genet*, 72(5):1117–1130, 2003.

[9] G. Barnett and C. Friedrich. Recent developments in ovarian cancer genetics. *Curr Opin Obstet Gynecol*, 16(1):79–85, 2004.

[10] M. Bella, R. Camisa, and S. Cascinu. Molecular profile and clinical variables in brca1-positive breast cancers. a population-based study. *Tumori*, 91:505–512, 2005.

[11] D. Daniel. Highlight: Brca1 and brca2 proteins in breast cancer. *Microsc Res Tech*, 59(1):68–83, 2002.

[12] E. Greenwood. Tumour suppressors: Unfold with brca1. *Nature Reviews Cancer*, 2(8), January 2002.

[13] T. Hernandez and S. Kambhampati. Integration of biological sources: current systems and challenges ahead. *SIGMOD Rec*, 33(3):51–60, 2004.

[14] R. Hoffmann and A. Valencia. A gene network for navigating the literature. *Nature Genetics*, 36:664, 2004. http://www.ihop-net.org/.

[15] D. Lewandowski. Web searching, search engines and information retrieval. *Information Services & Use*, 25(3), 2005.

[16] S. Li, Y. Ouyang, W. Wang, and B. Sun. Multi-document summarization using support vector regression. *Proceedings of Document Understanding Conference (DUC 07)*, 2007.

[17] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler. Yale: Rapid prototyping for complex data mining tasks. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-06)*, 2006. http://rapid-i.com/.

[18] W. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, pages 846–850, 1971.

[19] L. H. Reeve, H. Han, and A. D. Brooks. The use of domain-specific concepts in biomedical text summarization. *Information Processing and Management: an International Journal*, 2007.

[20] C. Stark, B. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, and M. Tyers. Biogrid: a general repository for interaction datasets. *Nucleic Acids Research*, 34, 2006. http://www.thebiogrid.org/.

[21] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. *Proceedings of KDD Workshop on Text Mining*, 2006.

[22] C.J. van Rijsbergen, S.E. Robertson, and M.F. Porter. New models in probabilistic information retrieval. *British Library Research and Development Report*, (5587), 1980. http://tartarus.org/~martin/PorterStemmer/.

[23] R. Verma, P. Chen, and W. Lu. A semantic free-text summarization system using ontology knowledge. *Proceedings of Document Understanding Conference (DUC 07)*, 2007.

[24] V. A. Yatsko and T. N. Vishnyakov. Familial dysautonomia and the expansion of the ashkenazi jewish carrier screening panel. *Genet Test*, 5(2), 2001.

[25] V. A. Yatsko and T. N. Vishnyakov. A method for evaluating modern systems of automatic text summarization. *Automatic Documentation and Mathematical Linguistics*, 41(3):93–103, 2007.

[26] I. Yoo and X. Hu. Empirical and theoretical comparisons of selected criterion functions for document clustering. *Machine Learning*, 55:365–371, 2004.

[27] I. Yoo and X. Hu. A comprehensive comparison study of document clustering for a biomedical digital library medline. *Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, 2006.

[28] I. Yoo, X. Hu, and I. Song. Integrating biomedical literature clustering and summarization approaches using biomedical ontology. *Proceedings of the 1st international workshop on Text mining in bioinformatics*, 2006.