BioSumm: A novel summarizer oriented to biological information

Alessandro Fiori

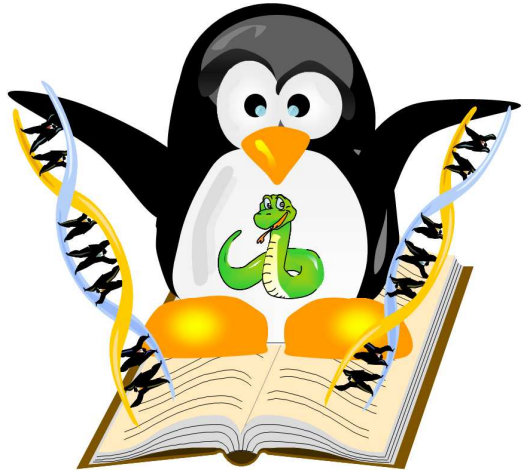
alessandro.fiori@polito.it

Dipartimento di Automatica e Informatica

Politecnico di Torino

Motivation

- The growing availability of large document collections has stressed the need of an effective management
- Most information is in free text not in structured data



Introduction

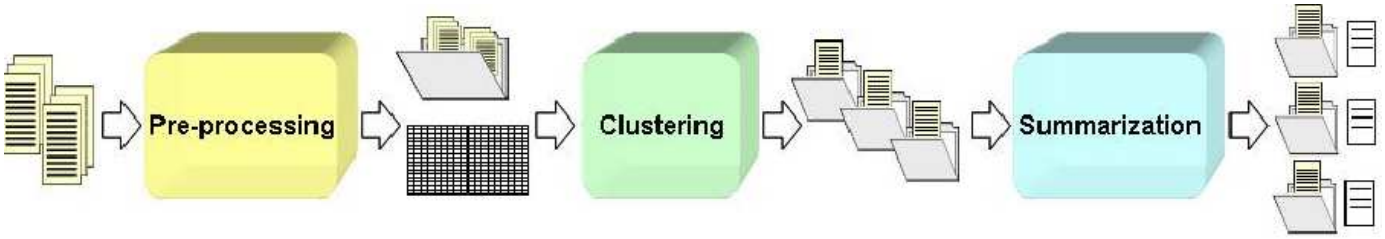
BioSumm

- Manages **unstructured** information contained in publicly available text collections
- Tailored for the domain of **biology** and for knowledge inference and biological validation
- Based on a novel automatic **text summarization** approach

Framework description

BioSumm framework

Modular architecture composed by three blocks



Preprocessing Block

Parses the inputs and represents them in the **Vector Space Model** (“Bag of words”)

Preprocessing Block

Parses the inputs and represents them in the **Vector Space Model** (“Bag of words”)

Inputs

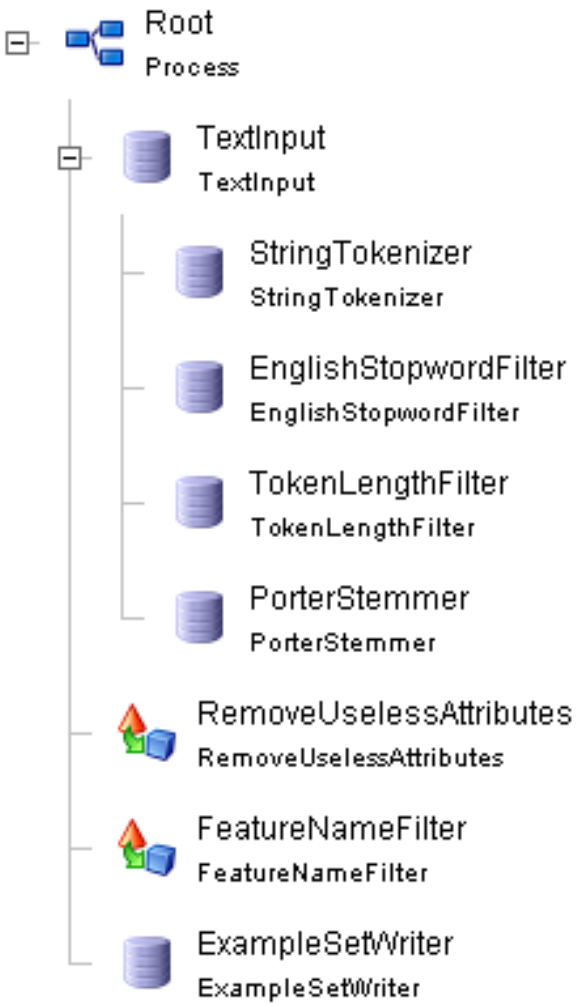
- Documents in the *PubMed Central* collection

Outputs

- “Cleaned” version of the document (*text only*)
- Matricial representation of the collection
- Each row is a *document*
- Each column is a *stem*
- Each cell contains the *tf-idf* of the word

Preprocessing Block

- Data Cleaning
  - Parses the *Pubmed Central* .xml files
  - Produces a purely textual output
- Matricial Representation
  - Exploits *RapidMiner* Text Plug-in
  - Customized work flow for BioSumm goals



Clustering Block

Divides unclassified and rather diverse texts into homogeneous clusters

Clustering Block

Divides unclassified and rather diverse texts into homogeneous clusters

Inputs

- “Cleaned” texts
- Matricial representation of the collection
- The number k of desired clusters

Outputs

- k clusters, homogeneous in terms of topics

Clustering Block

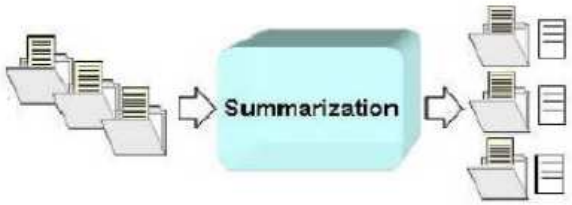
- Clustering performed by the CLUTO software package
- Clustering characteristics

  Method : Repeated Bisecting k-means

  Similarity Measure : Cosine

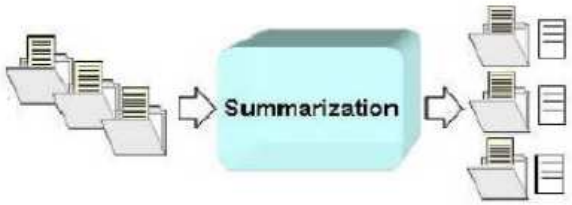
  Objective Function : $max \sum_{i=1}^{k} \sqrt{\sum_{v,u \in P_i} Sim(v,u)}$
- CLUTO is configured to minimize computational time
- Cluster quality is appropriate for BioSumm needs

Summarization Block



Produces summaries with the sentences useful for **knowledge inference** and **biological validation**

Summarization Block



Produces summaries with the sentences useful for **knowledge inference** and **biological validation**

Inputs

- The texts divided in clusters

Outputs

- A summary for each cluster

Summarization Block

- Has the advantages of traditional summarizers
- Overcomes their limitations in inferring knowledge of gene/protein relationships

Summarization Block

- Has the advantages of traditional summarizers
- Overcomes their limitations in inferring knowledge of gene/protein relationships
- The grading function biases sentence selection using a domain specific dictionary

$$\Gamma_j = \delta_j \cdot \sum_{k \in K} \omega_k \cdot \varphi_k$$

Summarization Block

- Has the advantages of traditional summarizers
- Overcomes their limitations in inferring knowledge of gene/protein relationships
- The grading function biases sentence selection using a domain specific dictionary

$$\Gamma_j = \delta_j \cdot \sum_{k \in K} \omega_k \cdot \varphi_k$$

- δ_j weights the number of occurrences of dictionary terms in sentence j and document i

Summarization Block

- Has the advantages of traditional summarizers
- Overcomes their limitations in inferring knowledge of gene/protein relationships
- The grading function biases sentence selection using a domain specific dictionary

$$\Gamma_j = \delta_j \cdot \sum_{k \in K} \omega_k \cdot \varphi_k$$

- δ_j weights the number of occurrences of dictionary terms in sentence j and document i
- ω_k is the number of occurrences, in document i, of a non stopword term k

Summarization Block

- Has the advantages of traditional summarizers
- Overcomes their limitations in inferring knowledge of gene/protein relationships
- The grading function biases sentence selection using a domain specific dictionary

$$\Gamma_j = \delta_j \cdot \sum_{k \in K} \omega_k \cdot \varphi_k$$

- δ_j weights the number of occurrences of dictionary terms in sentence j and document i
- ω_k is the number of occurrences, in document i, of a non stopword term k
- φ_k is the key words factor of Edmundson method

Summarization Block

$$\delta_j = \begin{cases} 1 & \text{if } \omega_g = 0, \\ & \forall g \in G \\ \alpha + \beta \cdot \sum_{g \in G} \hat{\omega}_g + \gamma \cdot \sum_{g \in G} (\omega_g - 1) & \text{otherwise} \end{cases}$$

Summarization Block

$$\delta_j = \begin{cases} 1 & \text{if } \omega_g = 0, \\ & \forall g \in G \\ \alpha + \beta \cdot \sum_{g \in G} \hat{\omega}_g + \gamma \cdot \sum_{g \in G} (\omega_g - 1) & \text{otherwise} \end{cases}$$

- α favors sentences that contain dictionary terms disregarding their number

Summarization Block

$$\delta_j = \begin{cases} 1 & \text{if } \omega_g = 0, \\ & \forall g \in G \\ \alpha + \beta \cdot \sum_{g \in G} \hat{\omega}_g + \gamma \cdot \sum_{g \in G} (\omega_g - 1) & \text{otherwise} \end{cases}$$

- α favors sentences that contain dictionary terms disregarding their number
- β weights the number of distinct dictionary term g

Summarization Block

$$\delta_j = \begin{cases} 1 & \text{if } \omega_g = 0, \\ & \forall g \in G \\ \alpha + \beta \cdot \sum_{g \in G} \hat{\omega}_g + \gamma \cdot \sum_{g \in G} (\omega_g - 1) & \text{otherwise} \end{cases}$$

- α favors sentences that contain dictionary terms disregarding their number
- β weights the number of distinct dictionary term g
- ω_g is the number of distinct occurrences of dictionary term g

Summarization Block

$$\delta_j = \begin{cases} 1 & \text{if } \omega_g = 0, \\ & \forall g \in G \\ \alpha + \beta \cdot \sum_{g \in G} \hat{\omega}_g + \gamma \cdot \sum_{g \in G} (\omega_g - 1) & \text{otherwise} \end{cases}$$

- α favors sentences that contain dictionary terms disregarding their number
- β weights the number of distinct dictionary term g
- ω_g is the number of distinct occurrences of dictionary term g
- γ weights the repetitions of dictionary term g

Summarization Block

$$\delta_j = \begin{cases} 1 & \text{if } \omega_g = 0, \\ & \forall g \in G \\ \alpha + \beta \cdot \sum_{g \in G} \hat{\omega}_g + \gamma \cdot \sum_{g \in G} (\omega_g - 1) & \text{otherwise} \end{cases}$$

- α favors sentences that contain dictionary terms disregarding their number
- β weights the number of distinct dictionary term g
- ω_g is the number of distinct occurrences of dictionary term g
- γ weights the repetitions of dictionary term g
- $(\omega_g - 1)$ counts all occurrences of dictionary term g, duplicated included

Summarization Block

$$\delta_j = \begin{cases} 1 & \text{if } \omega_g = 0, \\ & \forall g \in G \\ \alpha + \beta \cdot \sum_{g \in G} \hat{\omega}_g + \gamma \cdot \sum_{g \in G} (\omega_g - 1) & \text{otherwise} \end{cases}$$

- α favors sentences that contain dictionary terms disregarding their number
- β weights the number of distinct dictionary term g
- ω_g is the number of distinct occurrences of dictionary term g
- γ weights the repetitions of dictionary term g
- $(\omega_g - 1)$ counts all occurrences of dictionary term g, duplicated included

α is in the range $[1, +\infty)$, β is int the range $[0, 1]$, γ is in the range $[0, \beta]$

Experimental results

Experimental results

The experiments evaluate:

- The parameters and quality of the summarizer
- The quality of the output clusters
- The performances of the BioSumm framework

Collection	Journal	Size	Keywords
Bioinformation	Bioinformation	160	NO
Breast_Cancer	Breast Cancer Research	467	YES
J_Key	Breast Cancer Research Arthritis Research and Therapy	927	YES
Crit_Care	Critical Care	1460	NO

BioSumm vs. traditional summarizers

	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	Initially there was evidence that a TtoC variant in the CYP17 gene (allele frequency about 0.4) played a role in serum oestrogen and progesterone levels, and was associated with an increased risk of advanced disease.	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	Conclusions Considering the reported higher frequency of BRCA1 and BRCA2 germline mutations related to breast and ovarian cancer among Ashkenazi women in different countries, the results presented in this study were interpreted as showing a relatively lower than expected breast cancer mortality pattern among Ashkenazi women in the studied Brazilian cities.	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.

BioSumm vs. traditional summarizers

4	Two studies have estimated that mutations in the BRCA1 and BRCA2 genes only account for approximately 15% of the excess familial risk of the disease, while the contribution of the other known breast cancer susceptibility genes (TP53, PTEN, CHK2 and ATM) is even smaller.	Furthermore, the tumor genes and their mutations also appear to be responsible for an important, but still debated proportion of male breast cancers.
5	We also compared and combined parameter estimates from our new analyses with those from our logistic regression analyses of data on unaffected women from the Washington study, and derived a simple algorithm to estimate the probability that an Ashkenazi Jewish woman carries one of the three ancestral mutations in BRCA1 and BRCA2.	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	Mutations in TP53, and possibly in the ATM and CHK2 genes, seem to confer moderately increased risks of breast cancer but might explain only a very small proportion of familial aggregation.	These mutations were already reported in the literature or in the Breast Cancer Information Core electronic database.

Parameter evaluation

How to define the “Golden” Summary?

- A dictionary term extractor
- A traditional summary

Parameter evaluation

How to define the “Golden” Summary?

- A dictionary term extractor
- A traditional summary

BioSumm must find a balance

Parameter evaluation

How to define the "Golden" Summary?

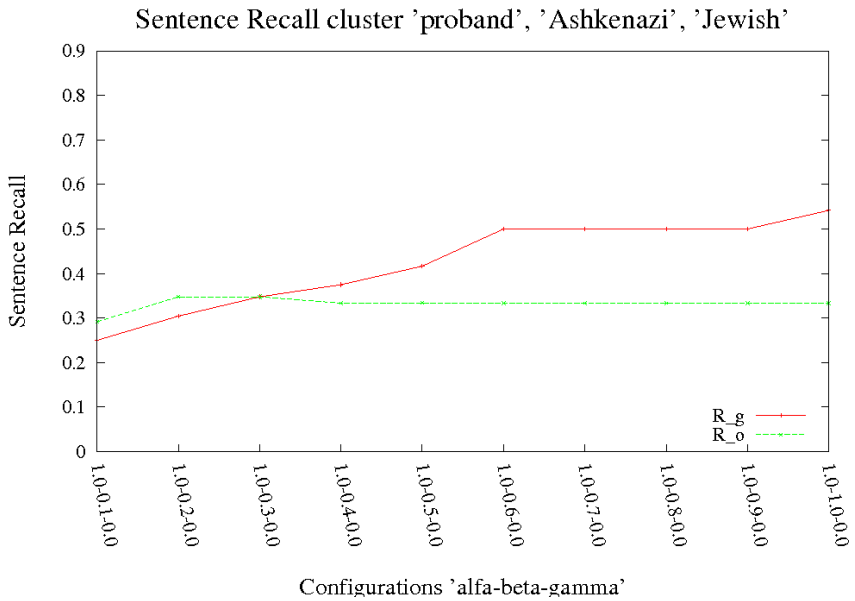
- A dictionary term extractor
- A traditional summary

BioSumm must find a balance

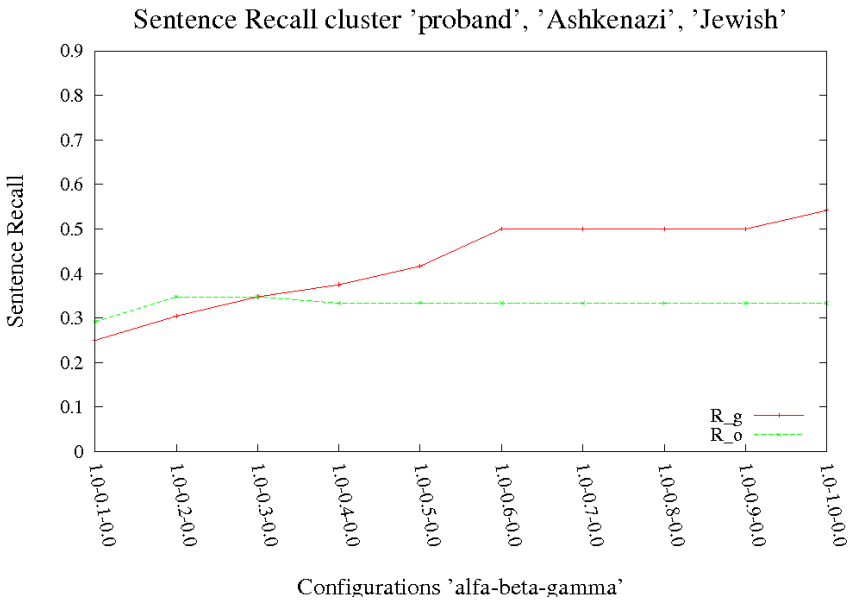
Evaluation Metrics

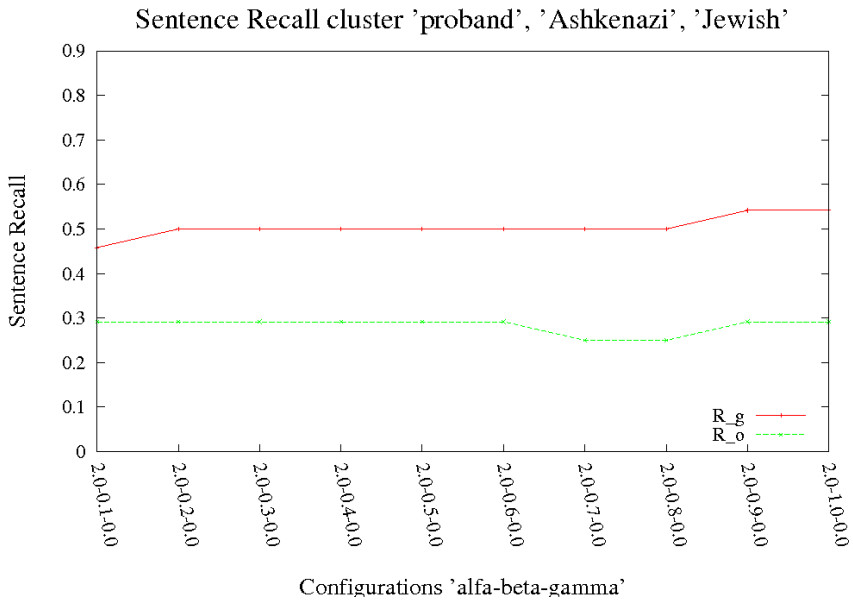
- Sentence Recall:
  - Recall_gene
  - Recall_ots
- Sentence Rank based Scores
  - Score_Gene: $S_g = \sum_{j \in O} \sum_g \hat{\omega}_g$
  - Score_ots: $S_o = \sum_{j \in O} s_j$

Parameter Evaluation

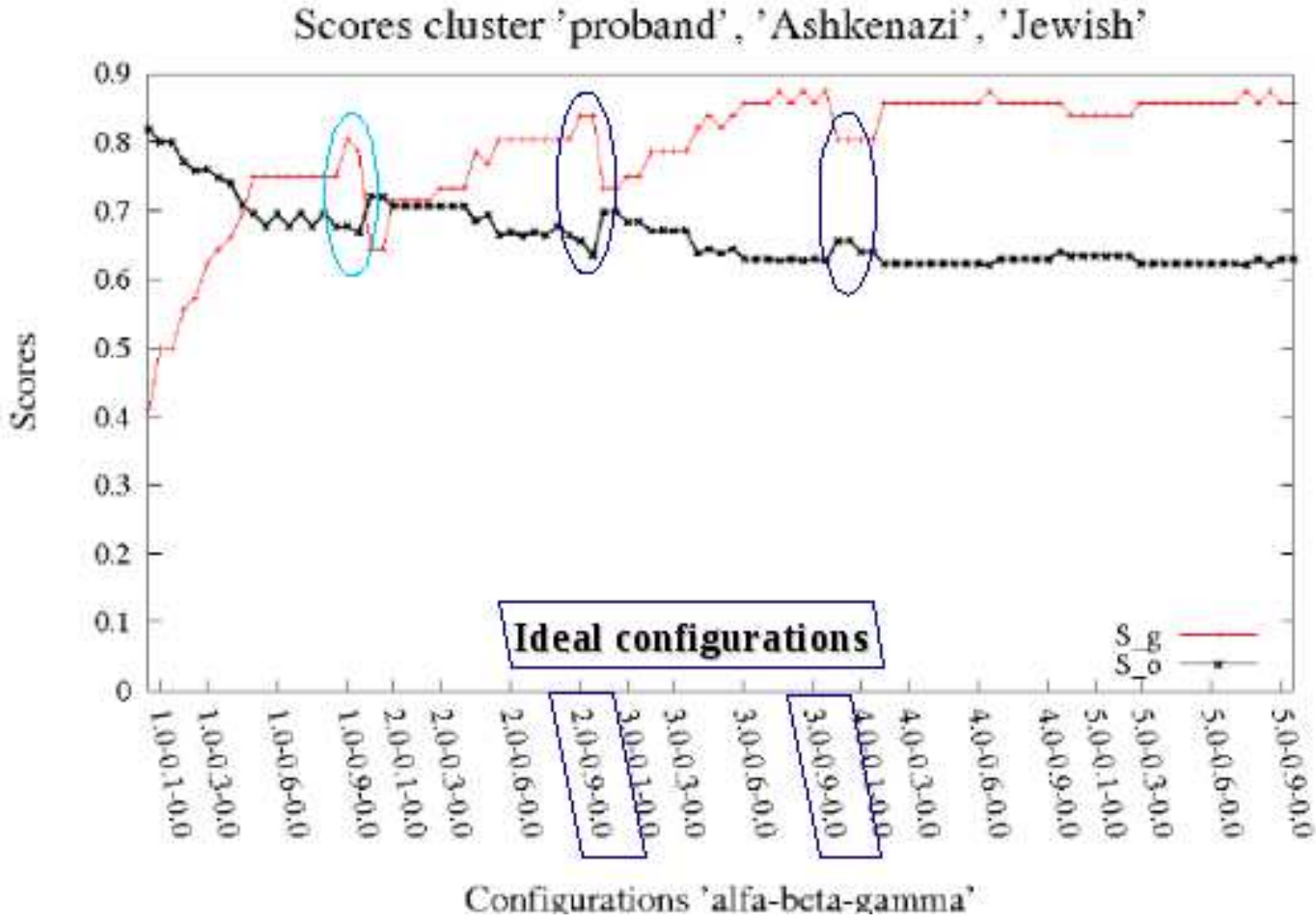


Parameter Evaluation





Parameter Evaluation



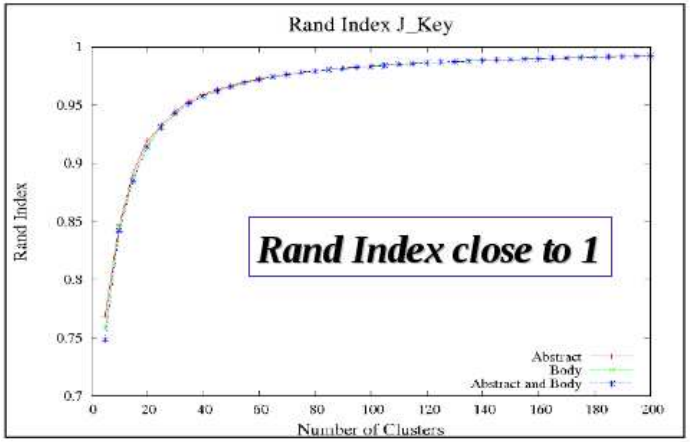
Other evaluations

- Clustering Quality
  - Rand Index:

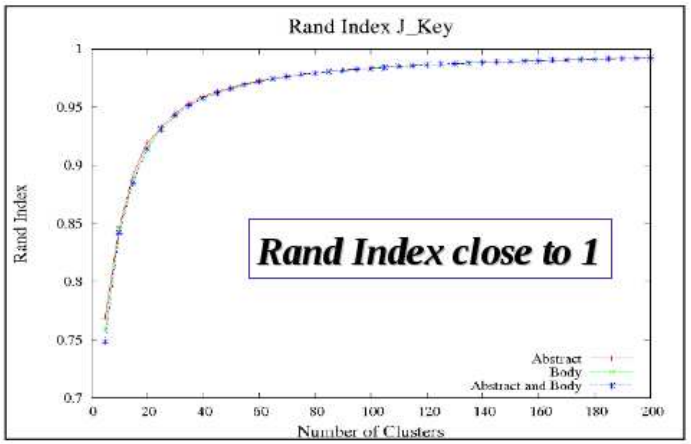
$$R(C,K) = \frac{a+b}{\binom{N}{2}}$$

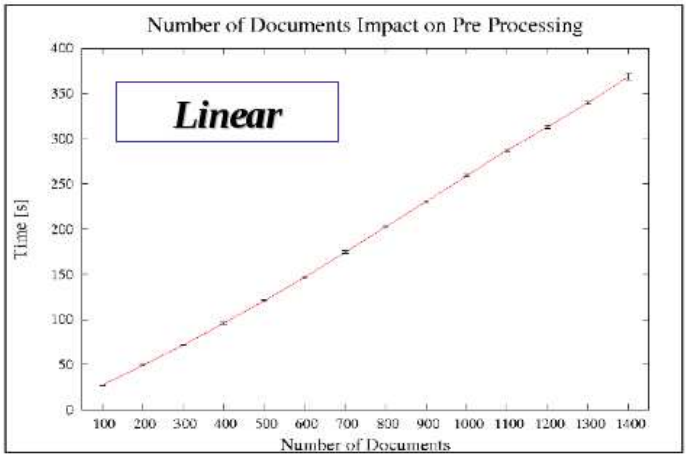
  - Measures the number of pairwise agreements between the BioSumm clusters and the clusters of the article Keywords
- Performance
  - The blocks have a roughly linear trend

Other evaluations

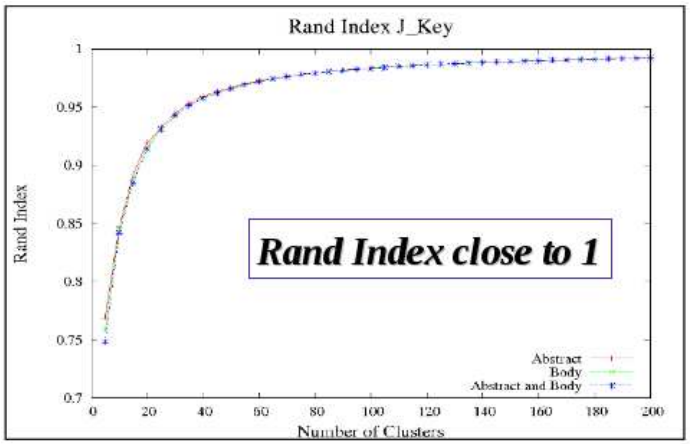


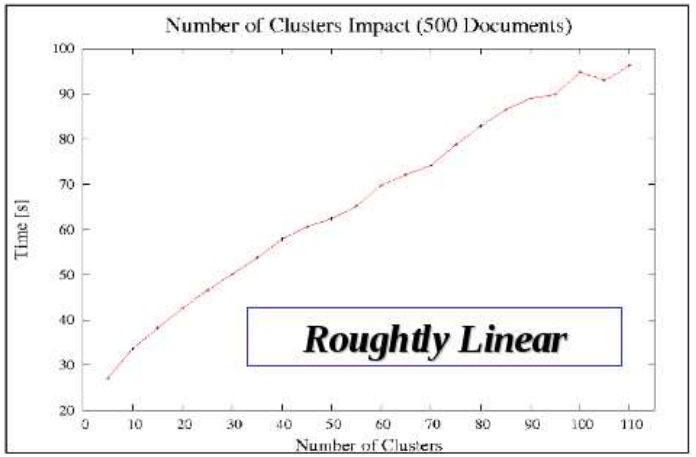
Other evaluations

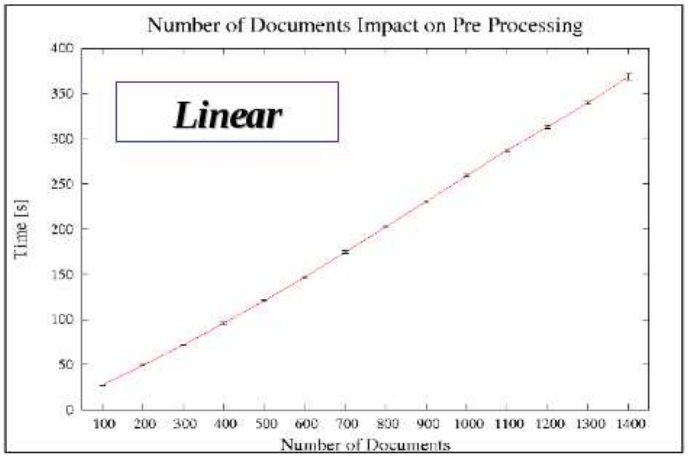




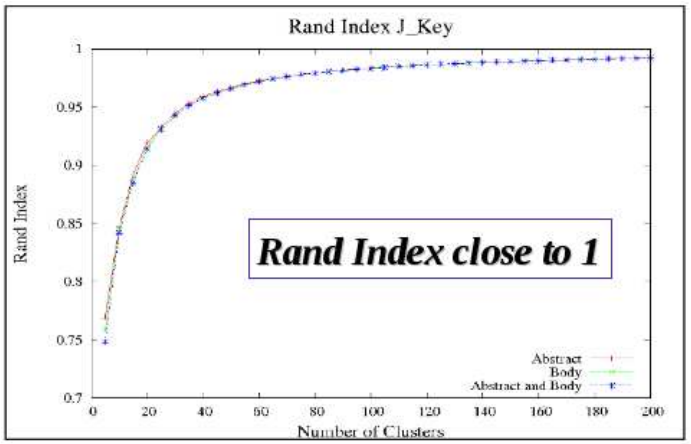
Other evaluations

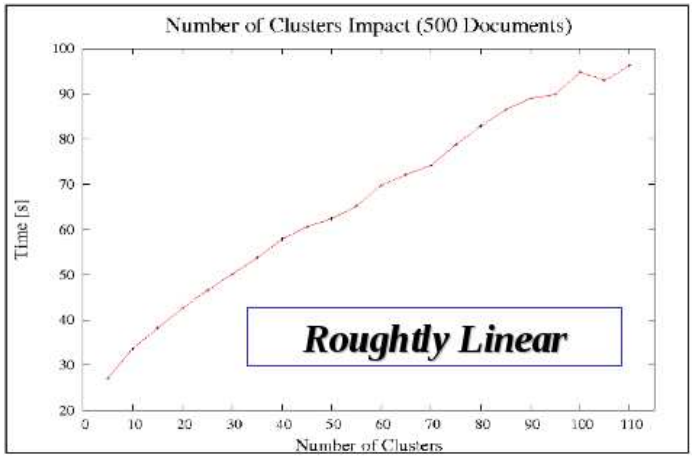


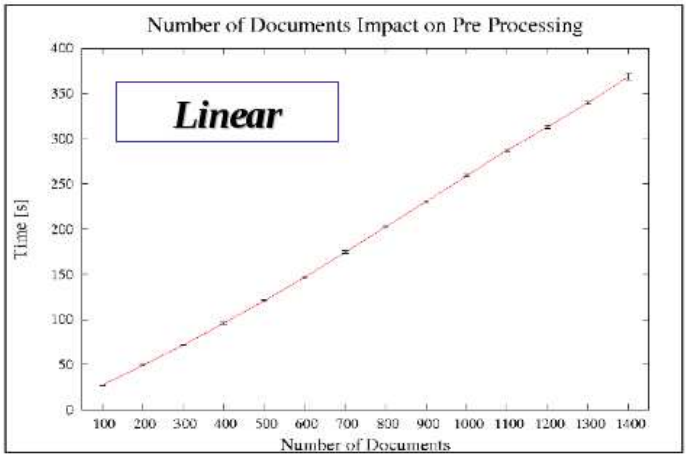


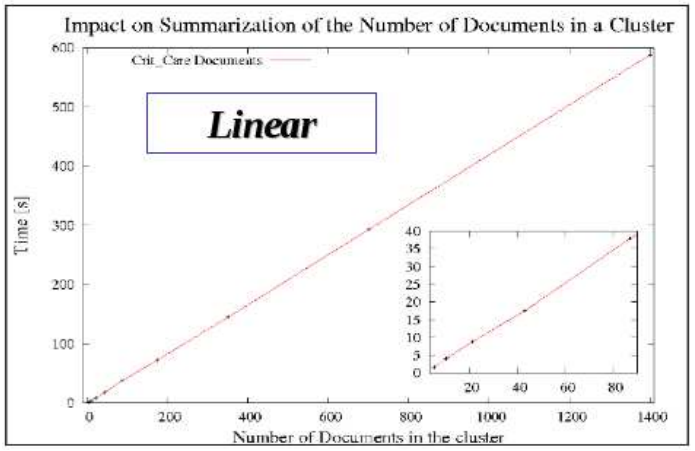


Other evaluations



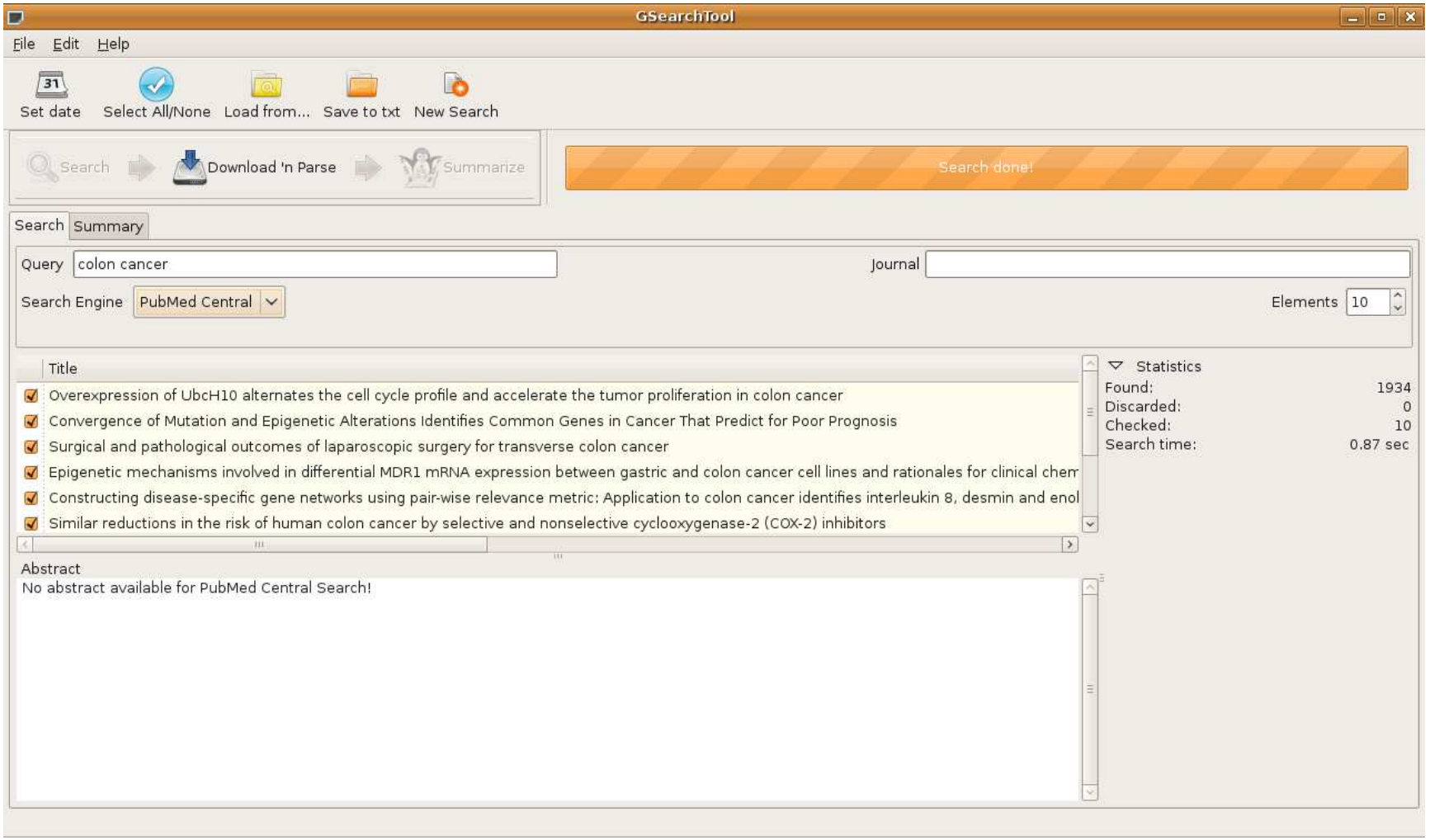




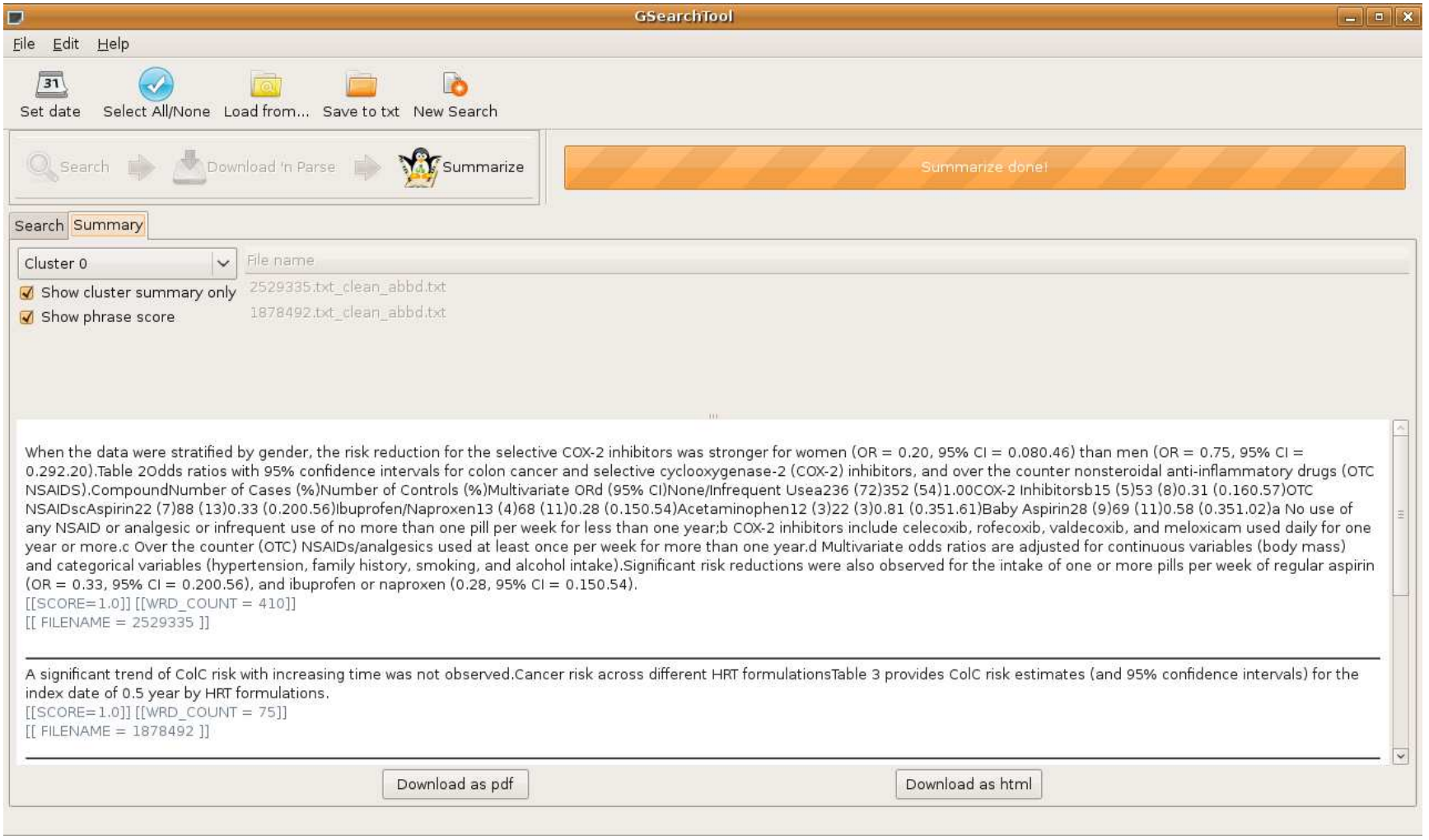


Program interface

Interface


GSearchTool
File Edit Help
Set date
Select All/None
Load from...
Save to txt
New Search
Search
Download 'n Parse
Summarize
Search done!
Search
Summary
Query
colon cancer
Journal
Search Engine
PubMed Central
Elements
10
Title
Overexpression of UbcH10 alternates the cell cycle profile and accelerate the tumor proliferation in colon cancer
Convergence of Mutation and Epigenetic Alterations Identifies Common Genes in Cancer That Predict for Poor Prognosis
Surgical and pathological outcomes of laparoscopic surgery for transverse colon cancer
Epigenetic mechanisms involved in differential MDR1 mRNA expression between gastric and colon cancer cell lines and rationales for clinical chem
Constructing disease-specific gene networks using pair-wise relevance metric: Application to colon cancer identifies interleukin 8, desmin and enol
Similar reductions in the risk of human colon cancer by selective and nonselective cyclooxygenase-2 (COX-2) inhibitors
Statistics
Found: 1934
Discarded: 0
Checked: 10
Search time: 0.87 sec
Abstract
No abstract available for PubMed Central Search!

Interface


GSearchTool
File Edit Help
Set date
Select All/None
Load from...
Save to txt
New Search
Search
Download 'n Parse
Summarize
Summarize done!
Search
Summary
Cluster 0
File name
Show cluster summary only
Show phrase score
2529335.txt_clean_abbd.txt
1878492.txt_clean_abbd.txt
When the data were stratified by gender, the risk reduction for the selective COX-2 inhibitors was stronger for women (OR = 0.20, 95% CI = 0.080.46) than men (OR = 0.75, 95% CI = 0.292.20).Table 2Odds ratios with 95% confidence intervals for colon cancer and selective cyclooxygenase-2 (COX-2) inhibitors, and over the counter nonsteroidal anti-inflammatory drugs (OTC NSAIDS).CompoundNumber of Cases (%)Number of Controls (%)Multivariate ORd (95% CI)None/Infrequent Usea236 (72)352 (54)1.00COX-2 Inhibitorsb15 (5)53 (8)0.31 (0.160.57)OTC NSAIDscAspirin22 (7)88 (13)0.33 (0.200.56)Ibuprofen/Naproxen13 (4)68 (11)0.28 (0.150.54)Acetaminophen12 (3)22 (3)0.81 (0.351.61)Baby Aspirin28 (9)69 (11)0.58 (0.351.02)a No use of any NSAID or analgesic or infrequent use of no more than one pill per week for less than one year;b COX-2 inhibitors include celecoxib, rofecoxib, valdecoxib, and meloxicam used daily for one year or more.c Over the counter (OTC) NSAIDs/analgesics used at least once per week for more than one year.d Multivariate odds ratios are adjusted for continuous variables (body mass) and categorical variables (hypertension, family history, smoking, and alcohol intake).Significant risk reductions were also observed for the intake of one or more pills per week of regular aspirin (OR = 0.33, 95% CI = 0.200.56), and ibuprofen or naproxen (0.28, 95% CI = 0.150.54).
[[SCORE=1.0]] [[WRD_COUNT = 410]]
[[FILENAME = 2529335]]
A significant trend of ColC risk with increasing time was not observed.Cancer risk across different HRT formulationsTable 3 provides ColC risk estimates (and 95% confidence intervals) for the index date of 0.5 year by HRT formulations.
[[SCORE=1.0]] [[WRD_COUNT = 75]]
[[FILENAME = 1878492]]
Download as pdf
Download as html

Future work

Future work

- Extend the BioSumm approach to other kind of summarizers (e.g., LSA based summarizers)
- Integrate ontology derived knowledge in the clustering phase
- Develop ad hoc clustering techniques
- Validate the effectiveness of the approach in other domains (e.g., financial articles)
- Analyze the gene/protein interactions extracted by the summarizer with UMLS ontology
- Graphical representation of biological concepts in the summary

Thanks for the attention!