2012 2nd International Conference on Biomedical Engineering and Technology
IPCBEE vol. 34 (2012) © (2012) IACSIT Press, Singapore

System for Biomedical Information Search in the Web

R. Guzmán Cabrera[1], J.A. Gordillo Sosa[2], A. González Parada[1], O.G. Ibarra Manzano[1] and M. Torres Cisneros[1]

[1] Universidad de Guanajuato
[2] Universidad Tecnológica del Suroeste de Guanajuato

Abstract. Due to the dramatic increase on available information on the Web, the users are in continuous demand of the appearance of new tools in order to find, filtrate and analyze the huge amount of data returned via search engines. In this paper we present a new tool for Web searching which starts from a small set of examples labeled by the user. From them a process of selection and weighting of words to form new queries ensures the recovery of more relevant documents. The validation of the implemented system was carried out by searches related to biomedicine, particularly related to cancer and its treatments.

Keywords: Biomedical, search, web

1. Introduction

One of the most dynamic and challenging environments nowadays is the Internet. Every day, the amount of data available increases dramatically, and, after a web query has taken effect, the huge number of results obtained make a detailed data analysis an impossible task to achieve. Therefore, we need to develop new tools to find, filter and analyze the files returned as a web query result.

According to a study about the size of the Web[1], Google claims to index more than 8 billion pages, MSN Beta about 5 billion pages, Yahoo! at least 4 billion and Ask/Teoma more than 2 billion. Two sources for tracking the growth of the Web are [6,7], although they are not kept up to date. Estimating the size of the whole Web is quite difficult, due to its dynamic nature. Nevertheless, it is possible to assess the size of the publically indexable Web. In Table 1 we show the summary of indexed pages by various search engines.

Table 1: Engines coverage %

	Engines Coverage %			
	Google	Msn	Teoma	Yahoo!
Coverage	76.3	62.03	57.58	69.28
	Engines Intersections %			
	Google	Msn	Teoma	Yahoo!
Google	-	55.8	35.56	55.63
Msn	78.4	-	49.56	67.38
Teoma	58.83	42.99	-	54.13
Yahoo!	67.96	49.33	45.21	-

The indexable Web [2] is defined as "the part of the Web which is considered for indexing by the major engines". In 1997, Bharat and Broder [3] estimated the size of Web indexed by Hotbot, Altavista, Excite and

[1] http://www.cs.uiowa.edu/~asignori/web-size/

Infoseek (the largest search engines at that time) at 200 million pages. They also pointed out that the estimated intersection of the indexes was less than 1.4 %, or about 2.2 million pages. Furthermore, in 1998, Lawrence and Giles [4] gave a lower bound of 800 million pages. These estimates have now become obsolete. The continuous increasing of the size of the Web shows an exponential behaviour. To be more acquainted with this, the number of registered networks in July 1999, were 56 million, 125 million in January 2001, and 172 million in January 2003. This situation has produced a growing need for tools that help people find, filter and analyze all these resources.

In Table 1, we show the number of results found by Google to certain requests in different years [1]. As an example, let us suppose that a certain user queries the Google search engine. The purpose of the queries is to find information about cancer as well as its cures and treatments. Table 1 shows the results produced by the search engine to the given queries. As we can see, even restricting the file type of the result, the number of results is so big that the user is unable to review all of them. This is because general purpose search engines deliver data related to many different domains. One solution to the problem is online query refinement [6], in which the user selects terms recommended by a helping tool.

Table 2: Google Request

Sample phrases	1998	2001	2003	2007	2012
Medical treatment	46,064	627,522	1,539,367	241,000,000	321,000,000
Prostate cancer	40,772	518,393	1,478,366	22,400,000	34,300,000
Vital organ	7,371	28,829	35,819	1,230,000	10,300,000

Table 3: Results about cancer

Sample phrases	Results	PDF
Cancer	175,000,000	65,400,000
Cure cancer	106,000,000	1,470,000
Cancer Bubi Cure	1,790,000	8,970

In this work, a tool for web searching is presented. Once interacting with it, the user must introduce a reduced amount of manually labeled documents that will be treated as a training set. Subsequently, every word of the mentioned set is selected and weighed, having the recovery of the most relevant documents related to the query as the goal in mind.

2. Web searching tool

In order to form the queries for searching on the Web, a set of manually labeled documents must be provided by the user, these documents will be used as the training set for our searching system.

In the first part, we begin by constructing a number of queries by combining the most significant words of each document; then, by using these queries, we perform a search on the Web for some additional training examples.

2.1. Query Construction

Before searching in the Web, it is necessary to determine the set of relevant words for each document in the training corpus.

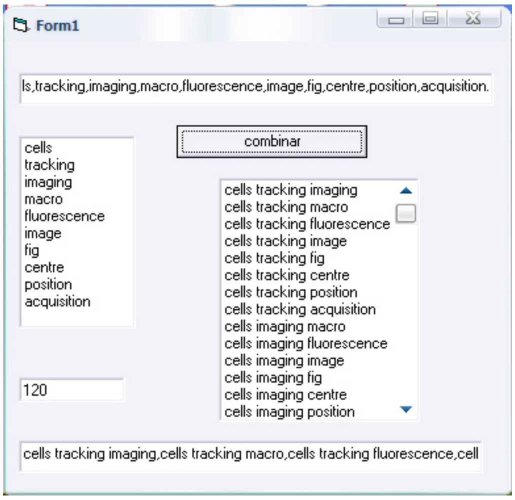
The criterion used for this purpose is based on a combination of the frequency of occurrence and the information gain of words. We consider that a word wi is relevant for document C if:

1. The frequency of occurrence of wi in C is greater than the average occurrence of all words (happening more than once) in that class. That is:
2. The information gain of wi with respect to C is positive.

2.2. Web Searching

The next step is using the generated queries to retrieve a set of relevant results from the Web. Based on the fact that most significant queries tend to retrieve the most relevant web pages, our method for searching the Web determines the number of downloaded examples per query in a direct proportion to its Γ-value. Therefore, given a set of M queries {q1,…, qM} for documents C, and considering that we want to download a total of N additional examples per document.

The user is able to select the number of documents to review, and the tool completes the task of delivering the most relevant results related to the query. In the process of completing the tests with this tool, a query about distributed generation was made with 1, 2, and 5 training sets. The results obtained reveal the accuracy and effectiveness of this method.



```
ls,tracking,imaging,macro,fluorescence,image,fig,centre,position,acquisition.
```

```
cells
tracking
imaging
macro
fluorescence
image
fig
centre
position
acquisition
```

```
120
```

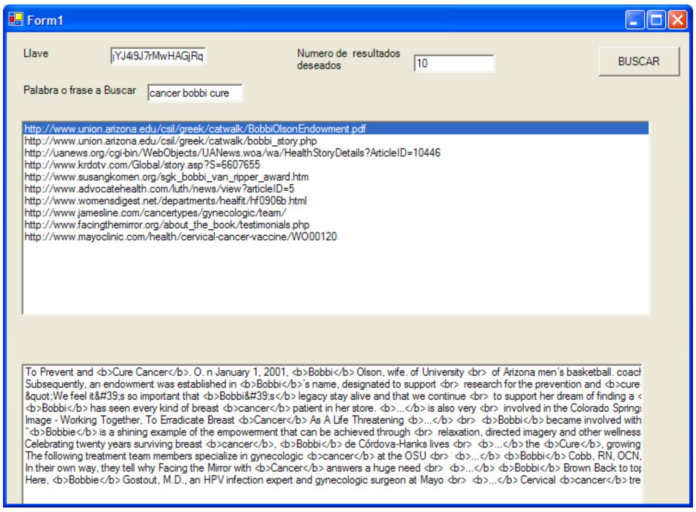
```
cells tracking imaging
cells tracking macro
cells tracking fluorescence
cells tracking image
cells tracking fig
cells tracking centre
cells tracking position
cells tracking acquisition
cells imaging macro
cells imaging fluorescence
cells imaging image
cells imaging fig
cells imaging centre
cells imaging position
```

```
cells tracking imaging,cells tracking macro,cells tracking fluorescence,cell
```

Fig. 1: Combination of the most frequent words of entire training set

Once the set of relevant words per class have been obtained, it is possible to construct the corresponding set of queries. Based on the method proposed by Zelikovitz and Kogan [5], we decided to construct queries of three words. In this way, we create as many queries per class as three-word combinations can be formed from its relevant words.

For the experiments, we started from a single training sample. We extracted the vocabulary and we selected the 10 words with the highest weight from it, according to our proposed method. Starting from those words, we proceed to the generation of the three-word combinations, as shown in figure 1.



Llave: |YJ4i9J7rMwHAGjRq

Numero de resultados deseados: 10

Palabra o frase a Buscar: cancer bobbi cure

```
http://www.union.arizona.edu/csil/greek/catwalk/BobbiOlsonEndowment.pdf
http://www.union.arizona.edu/csil/greek/catwalk/bobbi_story.php
http://uanews.org/cgi-bin/WebObjects/UANews.woa/wa/HealthStoryDetails?ArticleID=10446
http://www.krdotv.com/Global/story.asp?S=6607655
http://www.susangkomen.org/sgk_bobbi_van_ripper_award.htm
http://www.advocatehealth.com/luth/news/view?articleID=5
http://www.womensdigest.net/departments/healfit/hf0906b.html
http://www.jamesline.com/cancertypes/gynecologic/team/
http://www.facingthemirror.org/about_the_book/testimonials.php
http://www.mayoclinic.com/health/cervical-cancer-vaccine/WO00120
```

```
To Prevent and <b>Cure Cancer</b>. O. n January 1, 2001, <b>Bobbi</b> Olson, wife. of University <br>  of Arizona men's basketball. coach
Subsequently, an endowment was established in <b>Bobbi</b>'s name, designated to support <br>  research for the prevention and <b>cure
&quot;We feel it&#39;s so important that <b>Bobbi&#39;s</b> legacy stay alive and that we continue <br>  to support her dream of finding a <
<b>Bobbi</b> has seen every kind of breast <b>cancer</b> patient in her store. <b>...</b> is also very <br>  involved in the Colorado Spring
Image - Working Together, To Erradicate Breast <b>Cancer</b> As A Life Threatening <b>...</b> <br>  <b>Bobbi</b> became involved with
"<b>Bobbie</b> is a shining example of the empowerment that can be achieved through <br>  relaxation, directed imagery and other wellness
Celebrating twenty years surviving breast <b>cancer</b>, <b>Bobbi</b> de Córdova-Hanks lives <br>  <b>...</b> the <b>Cure</b>, growing
The following treatment team members specialize in gynecologic <b>cancer</b> at the OSU <br>  <b>...</b> <b>Bobbi</b> Cobb, RN, OCN,
In their own way, they tell why Facing the Mirror with <b>Cancer</b> answers a huge need <br>  <b>...</b> <b>Bobbi</b> Brown Back to to
Here, <b>Bobbie</b> Gostout, M.D., an HPV infection expert and gynecologic surgeon at Mayo <br>  <b>...</b> Cervical <b>cancer</b> tre
```

Fig. 2: Searching the Web

Figure 2 shows the 10 most relevant results. Now, the user must select more documents from these results for the training set.

The key idea is to teach the system to generate more queries, allowing with this the recovery of more relevant documents to the user, who has the choice of adding more documents to the training set in order to find new ones and so on.

Our approach is different from previous online query refinement tools in that it selects automatically the queries and the user only selects the new relevant documents to be added to the training set.

It is important to mention that the user must provide the number of desired results; as it can be seen in figure 2, the system shows the URL as well as a brief description, at the bottom of the window, and the content of the documents.

The number of results requested by the user is distributed among queries proportionally to the weight of each query. In this way, the user will have a greater number of results for a query with greater weight.

The system starts from a reduced set of labeled documents that are used as training set. The greater the number of samples provided for training, the better the queries that can be formed, which allows for better results from the Web.

3. References

[1] A. kilgarrif and G. Grefenstette, Introduction to the special issue on the web as corpus [Computational Linguistics 29(3): 333-348, 2003].

[2] E.Selberg, Towards Comprehensive Web Search [PhD thesis, University of Washington, 1999].

[3] S.Lawrence and C.L. Giles, Accessibility of information on the web [Nature 400:107-109, 1999].

[4] K.Bharat and A.Broder, A technique for measuring the relative size and overlap of public web search engines [WWW1998].