International Journal of Engineering Research and Development
***e**-ISSN: 2278-067X, **p**-ISSN: 2278-800X, www.ijerd.com*
Volume 3, Issue 11 (September 2012), PP. 20-25

Information represents about model based on Fingerprint’s text corpus

Carlos Balderas-Posada[2], Mario Rossainz-López[1], Yuridia Ramírez-Chocolatl[2], Mariela Alonso-Calpeño[2], Julio Zaldivar-López[2]
[1]*Benémerita Universidad Autónoma de Puebla*
[2]*Instituto Tecnológico Superior de Atlixco*

***Abstract*—The Information Retrieval (IR) area, is responsibility, mainly to the study of systems and techniques to assign index, search and give back valuable data to user, i.e. is used to investigate documents that exhibited a greater similar to the issued query [1]. In the research in IR, are currently in development models based on Fingerprint. In this research, we have developed different information to represent models based on Fingerprints, with the aim to show the documents/queries, so ensuring the management of large information’s volume efficiently. In the phase of experimentation, we took as a corpus to that granted by the TREC and they offered queries (with their respective Gold Standard), we had successful results with the evaluation program of the TREC we take it with a base line to the evaluation of our IR System.**

I. INTRODUCTION

The Information Retrieval Systems (*IRS*), have evolved from catalog’s automation to perform simple searches based on names or keywords, until the recent use of artificial intelligence techniques, to give it a view that allows you to select the relevant information [1,2].

Representation and indexation of information are arduous work in the analysis and design of an*IRS*, because, with these two techniques, the IRS depends to know how powerful could become it, not only in time response, also in the "quality" of the results obtained, given a set of queries.

There are already several models which allow representing information; this problem is not entirely resolved when speaking of very large documents collections, because practically the manipulation of information becomes impossible. This regard is currently developing information representation model based on Fingerprint, i.e. achievedtofind an unique and unambiguous way to represent a document through a chain of (much shorter) alphanumeric elements and on the other hand achieve to detect what part of the document can allows determining the topic of the document.

There are various information representation models on Fingerprints; one of the most used is presented by Schleimer [3], which proposes the Winnowing technique to generate the Fingerprint of a certain document. Harlistorm [4] developed a variant of the Winnowing algorithm, in which, after obtaining the Fingerprint of each document, all Fingerprints are input of a k-means algorithm (re-developed) where inside of it, it works with multisets (Fingerprints are taken as multisets) for to obtain a classification of documents. The N-Fingerprint algorithm developed in [5] is based on the creation of Fingerprints of documents according to the language of the corpus and n-grams, the Fingerprint DCT proposed in [6] use the fast Fourier transform for the creation of Fingerprint’s text, and finally, Benno Stein [7] presents a diffuse Fingerprint for the IR text-based.

The objective of this document is discuss of behavior that offers an SRI, which has been developed using various models for the generation of the fingerprint of each document and its indexing has made using the technique of posting list.

The rest of this document is structured in the following way, in section 2 presents an overview of the representation of information algorithms based on Fingerprints proposed for the management of documents/queries, under section 3 we detail how made to group similar documents according to the Fingerprint, under section 4 we reported the algorithm for the IR, in section 5 we describe the results that we obtained (until now) in the evaluation of our IRS and finally in section 6 we give a conclusion and future investigations for this project.

II. MODELS DEVELOPED TO REPRESENT INFORMATION

This is the general algorithm to obtain the fingerprint of a document:

```
functionobtain_fingeprints_docs( corpus )
begin
  while document = obtain_next_document( corpus ) do begin
    id_document   = obtain_id_document( document );
    text          = obtain_text_of_document( document ),
    string_fp = funcion_parser_text_to_fingerprint(text
    [, vocabulary_frecuecy_corpus|vocabulary_enumerated] );
    print_to_file( id_document, string_fp, “fp_documents_corpus.txt”);
  end
end
```

***Figure 1.*General algorithm of information representation based on Fingerprints**

In Figure 1 we have the function
funcion_parser_text_to_fingerprint(text[,vocabulary_frecuecy_corpus|vocabulary_enumerated]),
this function is the most important within the block of code, because it is responsible for the creation of the Fingerprint of the document.
For the development of Fingerprint of each document, we have developed the following models:

1. Based on the Karp-Rabin hash value of each term in the document: The Karp-Rabin hash value is a function that takes an *n*-gram (*n*-gram for us is a term) and parser it on a numerical value, based in calculatesthe value of the polynomial$H(c_1, \ldots, c_k) = c_1 * b^{k-1} + c_2 * b^{k-2} + \cdots + c_{k-1} * b + c_k$, where b is any basis (for our purpose of representation $b = 1$). $c_1, \ldots, c_k$is the n-gram to convert. The set of all values of all document's terms is the fingerprint of the same document.
2. Based on Fingerprint created by winnowing algorithm: We use the approach proposed in [3], we take part of the Winnowing algorithm for creating Fingerprints documents. Winnowing algorithm receives the value of $w = 25$, $n = 5$ and the value of $b = 1$ (b is used in the Karp-Rabin function, which was described above).
3. Based on the phonetic representation (Soundex for Spanish[1]) of each term that composes the document: This algorithm takes a term and get the phonetic chain of it, the phonetic code is dissimilar to the English language, because the Spanish language takes into account the double letters (ll and ch) and tildes (ñ). The Soundex algorithm for the Spanish language is in [8], but, this is designed in an Oracle runtime. For our work, we translated the runtime to the language objective of this work (*AWK*). The set of all phonetic codes of all document terms is the fingerprint of the same document.
4. Based on the sum of numeric values of the characters, it is, sum the ASCII values of each character (different) that compose each term in the document, for example: banana → ban → 98+97+110 → 305. The set of all values of all document's terms is the fingerprint of the same document.
5. Based on the enumeration of the vocabulary of the corpus: In this model, we obtain the corpus vocabulary to deal with, every term is arranged alphabetically and then, assign each a consecutive number. After having listed the vocabulary, each term in the corpus is replaced by its corresponding number.

III. DOCUMENT CLUSTERING MODEL

After obtaining a Fingerprint of each document in the corpus, with representation models presented above, we designed a method of grouping (clustering) documents with the same or similar Fingerprint, in order to reduce the size of the IRS indexing, both in space storage and search time to time of *IR*. In the clustering of Fingerprints criteria we used a measure the degree of similarity using the Jaccard Coefficient: $J(A,B) = \frac{|A \cap B|}{MIN(|A|,|B|)} \geq 0.5$ where A and B are Fingerprints of two documents. This measure is to verify the containment (perhaps all) of the largest document in the small document.

We create a single posting list, where it can be an entry for each document in the corpus, in the case that none has had a degree of similarity greater than 0.5, otherwise the entry has grouped of similar documents. We call it: the Indexing over posting list.

In the figure 1, we can see how fingerprint algorithm works.

[1] Because the corpus is in Spanish language.

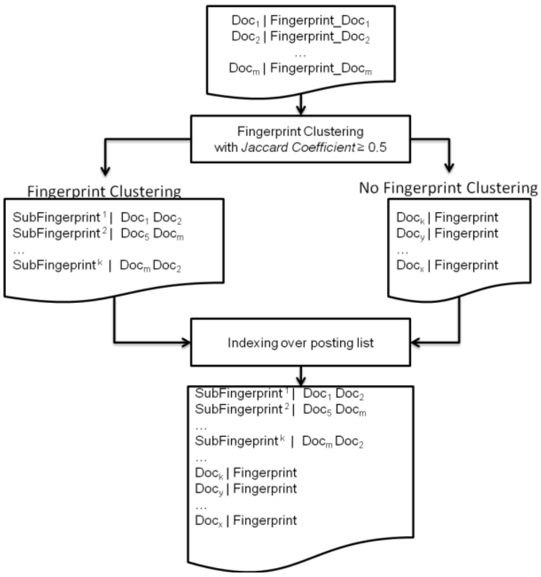

Doc$_1$ | Fingerprint_Doc$_1$
Doc$_2$ | Fingerprint_Doc$_2$
…
Doc$_m$ | Fingerprint_Doc$_m$
Fingerprint Clustering with *Jaccard Coefficient* ≥ 0.5
Fingerprint Clustering
No Fingerprint Clustering
SubFingerprint^1 | Doc$_1$ Doc$_2$
SubFingerprint^2 | Doc$_5$ Doc$_m$
…
SubFingeprint^k | Doc$_m$ Doc$_2$
Doc$_k$ | Fingerprint
Doc$_y$ | Fingerprint
…
Doc$_x$ | Fingerprint
Indexing over posting list
SubFingerprint^1 | Doc$_1$ Doc$_2$
SubFingerprint^2 | Doc$_5$ Doc$_m$
…
SubFingeprint^k | Doc$_m$ Doc$_2$
…
Doc$_k$ | Fingerprint
Doc$_y$ | Fingerprint
…
Doc$_x$ | Fingerprint

Figure 1. DFD's Fingerprint process

IV. INFORMATION RETRIEVAL MODEL

After obtain a single posting list, an algorithm was designed for the RI, that algorithm interacts with the indexing of the corpus and the representation of queries under the same Fingerprint scheme.

The propose of the algorithm is based on a combination of posting list management and evaluation of the similarity between the posting list entries and queries. The similarity is calculated using the similarity JaccardCoefficent, it was explained in Section 3. Identifiers of the documents returned by the algorithm are ranked from highest to lowest; we discard those with a lower level of similarity to 0.5.

Figure 2 shows the information retrieval model:

```
functioninformation_retrieval( all_queries,finally_posting_list )
begin

  queries = index_querys_in_hash( all_queries );

  h_posting_list = index_posting_list_in_hash( finally_posting_list );

  foreach x in queries do begin
    foreach y inh_posting_listdo begin
      jk = Function_Jaccard( queries [ x ], y );

      ifjk> 0 then begin
        hash_result_query[ h_posting_list[ y ] ] = jk;
      end
    end

    sort_result_max_min(hash_result_query, queries_sorted);

    foreach z inqueries_sorteddo begin
      print_results( x, queries[ x ], z, queries_sorted[ z ], "results.txt");
    end
  end

end
```

Figure 2. **General model of the IRS**

V. ANALYSIS OF RESULTS

For the evaluation of IRS design, we use a corpus of news in Spanish[2]. The news corpus is divided into 5 sub corpus, each one with its own set of queries and the Gold Standard of them.
The queries provided for the experiment were 10, which are shown in Table 1.

[2] http://trec.nist.gov/data/docs_noneng.html/

Table 1. Description of the queries used in the evaluation of the IRS

Subcorpus	ID query	Query	Relevant Documents
1	Q_10	México es importante país de tránsito en la guerra antinarcótica.	206
	Q_11	Derechos a las aguas de los ríos en la región fronteriza entre México y los Estados Unidos	105
2	Q_01	Oposiciónmexicana al TLC	211
	Q_03	Polución en el Distrito Federal de México	164
3	Q_14	El monopolio petrolero PEMEX tiene mucha influencia en México	281
	Q_15	La disputa sobre la pesca ha ocasionado la captura de barcos de pesca de los Estados Unidos	7
4	Q_04	El papel de México en la OEA	97
	Q_05	Maquiladoras en la economía mexicana	257
5	Q_24	Prevención de SIDA en México	131
	Q_25	Programa de privatización de empresas mexicanas	359

We validated the 5 types of indexing provided in this investigation over 5 subcorpus vs. the *TREC* Evaluation System[3] (*TREC-ES*).

Table 2 shows the results of the evaluation of our *IRS* vs. *TREC-ES,* based on subcorpus 1with their respective queries:

Table 2. Results for Subcorpus 1 vs. *TREC-ES*

	Information representation model used in the indexing of SRI				
Values offered by TREC-ES	**Karp-Rabin**	**Soundex Spanish**	**Add ASCII characters without repetition**	**Numbering of the vocabulary**	**Winnowing**
num_q	2	2	2	2	2
num_ret	1590	1181	1576	241	1535
num_rel	311	311	311	311	311
num_rel_ret	287	277	271	73	260
MAP[4]	0.2483	0.3373	0.2116	0.0863	0.144
gm_ap	0.2464	0.3365	0.2114	0.0793	0.1366
R-prec	0.2576	0.3459	0.2595	0.2122	0.1686
bpref	0.7687	0.7615	0.6852	0.2009	0.5854
recip_rank	0.2667	0.75	0.6	1	0.5357
P5	0.3	0.5	0.4	0.2	0.1
P10	0.45	0.6	0.25	0.45	0.05
P15	0.4667	0.5667	0.2	0.4667	0.2
P20	0.45	0.5	0.2	0.45	0.25
P30	0.4	0.5167	0.25	0.4	0.2
P100	0.29	0.415	0.275	0.27	0.175
P200	0.265	0.315	0.2275	0.1825	0.16
P500	0.214	0.242	0.19	0.073	0.166
P1000	0.1435	0.1385	0.1355	0.0365	0.13

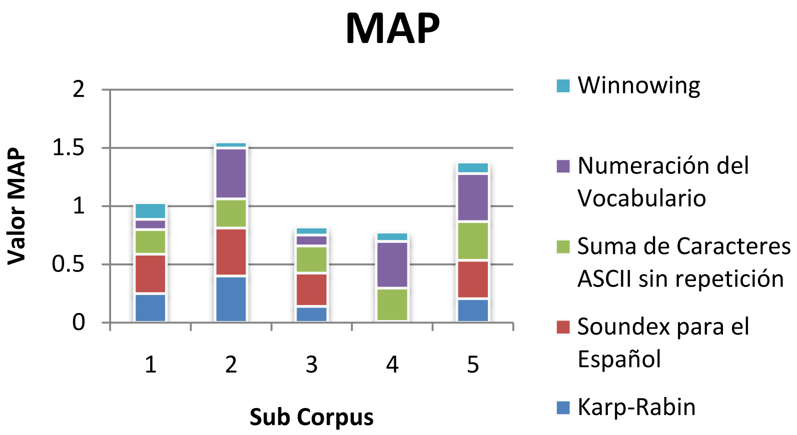
Although the model of representation using Karp-Rabin returned 287 relevant documents of the 311, its *MAP* is less than that offered by the Soundex for Spanish; this is because the documents retrieved by using Soundex version have

[3] http://trec.nist.gov/data/reljudge_noneng.html

[4] *Medium Average Precision (MAP).*

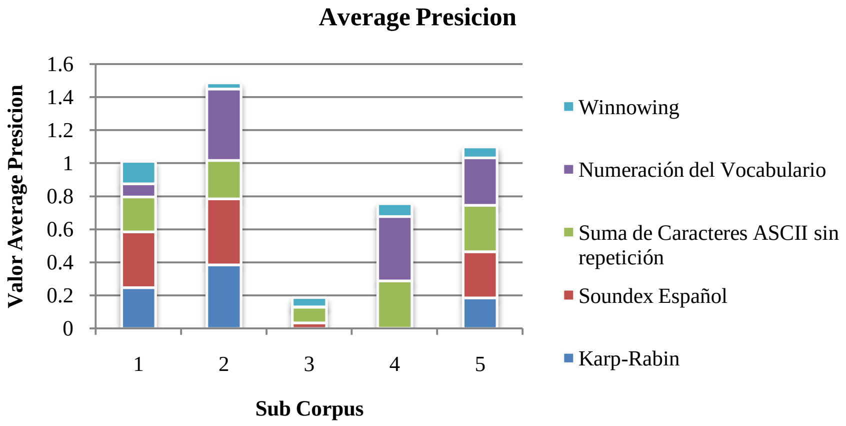
better ranked than Karp-Rabin. From the results obtained by analyzing the parameters *Average Precision*, *R-precision*, *bpref* and *recip_rank*, Soundex show better values than the others models of representation.

Another aspect to note is that the first 30 retrieved documents show an accuracy above 50%, this is an important aspect, because the users of an IRS, always looking for your information in the top 50 (or less) documents returned.
Graph 1 shows clearly proposed before:



Graph 1.MAP results offered by *TREC-ES* with subcorpus 1, 2, 3, 4 and 5

The Graph 2 shows Average Precision, we can see the same result of last Graph 1, Numering of the Vocabulary, Soundex and Add ASCII characters without repetition are the best models of Fingerprints.

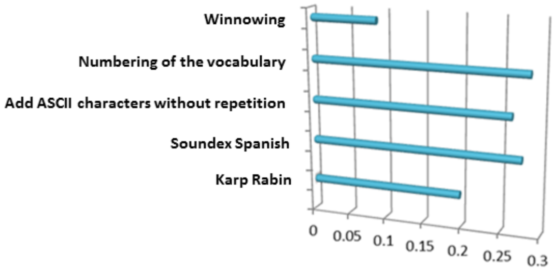


Graph 2. Average Precision results of own IRS, with the different Fingerprint.

Table 3 and Graph 5 show the average *MAP* values obtained in 5 subcorpus. It is important to note that indexing using the representation model numbering of vocabulary allows that the *IRS* provide many documents relevant to different queries, independent of the corpus on which it is working, considering the measures offered by the *TREC-ES* in the framework of *TREC*, for the task of ad hoc retrieval.

Table 1. Table with the average values of *MAP*

	Average MAP
Karp Rabin	0.19746
Soundex Spanish	0.27514
Add ASCII characters without repetition	0.26202
Numbering of the vocabulary	0.28372
Winnowing	0.08846



Graph 5. Average MAP with 5 subcorpus.

VI. CONCLUSIONS

The overall objective of the work is completed, developing several algorithms to represent large volumes of Fingerprint-based information. We validated using the *TREC*. The Information Retrieval system, indexes the documents generated by the algorithm Firgerprints Numbering Vocabulary, and that this representation was the one that provided better accuracy. The second best performing algorithm was the Soundex algorithm for the Spanish.

The advantages of the model representation of information are:

- Reduce storage space considerably, because it occupies smaller alphanumeric strings.
- Allowed to spend less time when comparing strings smaller than the original strings in the document.
- The information stored within the posting list is less and they prevent the RAM is 100% loaded at runtime. This involves the grouping of Fingerprints in conjunction with the above.
- The grouping and RI models emphasize the problem A $\subseteq/\supseteq$ B, which in classical *IR* models do not take into account.
- Greater efficiency in time compared with techniques using vector representation *RI*.
- According to the results of the evaluation, the final representation used to be based within the search engine (Numbering of vocabulary) provides excellent results for the average user in the top 30 relevant documents.

Recommendations to continue this research project, we propose the following points:

- Test our models in other corpus with differentlanguages, to verify if they are efficient.
- Migrating to programming languages that provide a wider range of representation of integers (for example, Java C++ or C # with BigInteger class, or minimally provide the use of 64-bit integers).
- Test our models with a restricted domain corpus to re-validate the clustering of Fingerprints.
- Re-implement the models using tree representation techniques to search text patterns in Fingerprints.

REFERENCES

[1]. Baeza-Yatez, R.; et al. *Modern Information Retrieval*. Inglaterra: Addison Wesley, 1999.

[2]. Grossman, D.; et al. *Information Retrieval and Heuristics*. Paises Bajos: Springer, 2004.

[3]. Schleimer, Saúl; et al. Winnowing: Local Algorithms for Document Fingerprinting. SIGMOD 2003. EUA: ACM Press, 2003.

[4]. Hamid, O. A.; et al. *Detecting the Origen of Text Segments Efficiently*. Proceedings of the 18th International Conference on WWW. EUA: ACM Press, 2009.

[5]. Parapar, Javier; et al. Evaluation of Text Clustering Algorithms with N-Gram-Based Document Fingerprint. Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval. Francia: Springer, 2009.

[6]. Seo, Jangwon; et al. *Local Text Reuse Detection*. Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. EUA: Springer, 2008.

[7]. Stein, Benno. *Fuzzy-Fingerprint for Text-Based Information Retrieval*. Proceedings of I-Know '05. Austria: Maurer Tochtermann, 2005.

[8]. Runtime de Oracle Soundexpara el Español [online]. http://oraclenotepad.blogspot.com/2008/03/soundex-en-espaol.html. [searcher: 28th March 2011].