SARIPOD: A SYSTEM BASED ON HIERARCHICAL SMALL WORLDS AND POSSIBILISTIC NETWORKS FOR INTERNET INFORMATION RETRIEVAL

Bilel Elayeb*, Montaceur Zaghdoud*, Mohamed Ben Ahmed*
** RIADI-GDL laboratory, The National School of Computer Sciences (ENSI), Manouba University, 2010 Tunisia.*
{Bilel.Elayeb, Mohamed.Benahmed}@riadi.rnu.tn, Montaceur.Zaghdoud@ensi.rnu.tn

Fabrice Evrard**
*** The Computer Science Research Institute of Toulouse (IRIT), 02 Chamichel Street, 31071 Toulouse Cedex 7 French.*
Fabrice.Evrard@enseeiht.fr

ABSTRACT

This paper presents an Internet information retrieval system based on Hierarchical Small-Worlds (HSW) and Possibilistic Networks (PN). The first HSW, for the words of the French language, is used to take account of the dependences between these words. The second HSW is devoted to the web pages required and translated in the same way the dependences between these pages. The PN generates the mixing of these two HSW in order to organize the sershed documents according to the user profile. Our system propose a mixed approach for Information Retrieval Model based on possibility and necessity measures. This model encodes relationship dependencies existing between query terms and web documents through naïve possibilistic networks and quantifies these relationships by two measures: possibility and necessity. The retrieved documents are those which are necessarily or possibly relevant given a user's query. The search process restores the plausibly or necessarily relevant documents for a user need.

KEYWORDS

Information Retrieval, Hierarchical Small-Worlds, Possibilistic Networks, Strategie crawl, possibilistic relevance.

1. INTRODUCTION

The key issue of Information Retrieval (IR) is that documents must be retrieved from a large document collection in response to a user's need, often on the basis of poor information. Known models in the literature (Boolean, vector space, probabilistic, Bayesian) represent documents and queries only through weighted lists of terms and a measure of relevance is computed (vector space similarity, probabilistic relevance) based on those weighted lists. Devising a proper weighting scheme seems to be the fundamental element of actual IR models since the computation of relevance relies on it. Usually, the weighting scheme is the result of several combinations: term frequencies in document (*tf*), term frequencies in the whole collection (*idf*) and document length (*dl*). Whatever the used model, the response to a user need is a list of documents ranked according to a unique relevance value. Many approaches consider term weights as probability of relevance. In such models the incompleteness of information is not considered when representing or evaluating documents given a query. Yet, the rough nature of document descriptions (a multiset of terms) and of the query description (a list of terms) are hardly compatible with the high precision of relevance values obtained by current methods.

The aim of this paper is to propose basic steps towards an IR mixed approach based on possibility and necessity measures. Instead of using a unique relevance value, we propose a possibilistic approach for computing relevance. This model should be able to infer propositions like:

- It is plausible to a certain degree that the document is relevant for the user need.
- It is almost certain (in possibilistic sense) that the document is relevant to the query.
- Document d_1 is more appropriate than document d_2 or the set $\{d_1, d_2\}$ is better than the set $\{d_3, d_4\}$.

The first kind of proposition is meant to eliminate irrelevant documents (weak plausibility). The second answer focuses attention on what looks very relevant. The third proposition suggests that, since the raw

information on documents is more qualitative than quantitative, ordinal approaches to the problem may be interesting as well. The use of probability theory in the definition of relevance given a query does not account for our limited knowledge of the relevance of a document, since it does not consider imprecision and vagueness intrinsic to relevance (Brini et Boughanem, 2003).
We present here the hierarchical small world graph in section 2 and we briefly recall some notions of possibility theory, in section 3. We describe along section 4, the general architecture of SARIPOD system and we present the fonctionnality of each module with some examples. Section 5 suggests future works.

2. HIERARCHICAL SMALL-WORLDS

Recent work in graph theory has revealed a set of features shared by many graphs observed "in the field" These features define the class of "hierarchical small-worlds" networks (Watts and Strogatz, 1998). The relevant features of a graph in this respect are the following:

D : the density of the network. HSWs typically have a low D, i.e. they have rather few edges compared to their number of vertices.
L : the average shortest path between two nodes. It is also low in a HSW.
C : the clustering rate. This is a measure of how often neighbors of a vertex are also connected in the graph. In a HSW, this feature is typically high.
I : the distribution of incidence degrees (i.e. the number of neighbors) of vertices according to the frequency of nodes (how many nodes are there that have an incidence degree of 1, 2, ..., n). In a HSW network, this distribution follows a power law.
As a mean of comparison, table 1 (Gaume et al., 2004) shows the differences between random graphs (nodes are given, edges are drawn randomly between nodes), regular graphs and HSWs. The graph of the web belongs to the class of HSW (Douglas et Houseman, 2002).

Table 1. Comparing classes of graphs

with equal D	L	C	I
Random graphs	small L (short ways)	small C (not of aggregates)	Poisson Law
HSW	**small L (short ways)**	**High C (the aggregates)**	**power law**
Regular graphs	High L (long ways)	High C (the aggregates)	constant

3. POSSIBILISTIC LOGIC

Possibility theory introduced by Zadeh (Zadeh, 1978) and developed by Dubois and Prade (Dubois et Prade, 1987), handles uncertainty in the interval [0,1] called possibility scale, in a qualitative or quantitative way.

3.1 Possibility distribution

Possibility theory is based on possibility distributions. The latter, denoted by π, are mappings from Ω (the universe of discourse) to the scale [0,1] encoding partial knowledge on the world. The possibility scale is interpreted in two ways. In the ordinal case, possibility values only reflect an ordering between possible states; in the numerical scale, possibility values often account for upper probability bounds.

3.2 Possibility and necessity measures

A possibility distribution π on Ω enables events to be qualified in terms of their plausibility and their certainty, in terms of possibility and necessity measures respectively (Dubois et Prade, 1987).
The possibility $\Pi(\mathrm{A}) = \max_{x \in A} \pi(x)$ of an event A relies on the most normal situation in which A is true.

The necessity $N(A) = \min_{x \notin A} 1 - \pi(x) = 1 - \Pi(\neg A)$ of an event A reflects the most normal situation in which A is false.

The width of the gap between N(A) and $\Pi(A)$ evaluates the amount of ignorance about A. Note that $N(A) > 0$ implies $\Pi(A) = 1$. When A is a fuzzy set this property no longer holds but the inequality $N(A) \leq \Pi(A)$ remains valid.

3.3 Possibilistic Networks (PN)

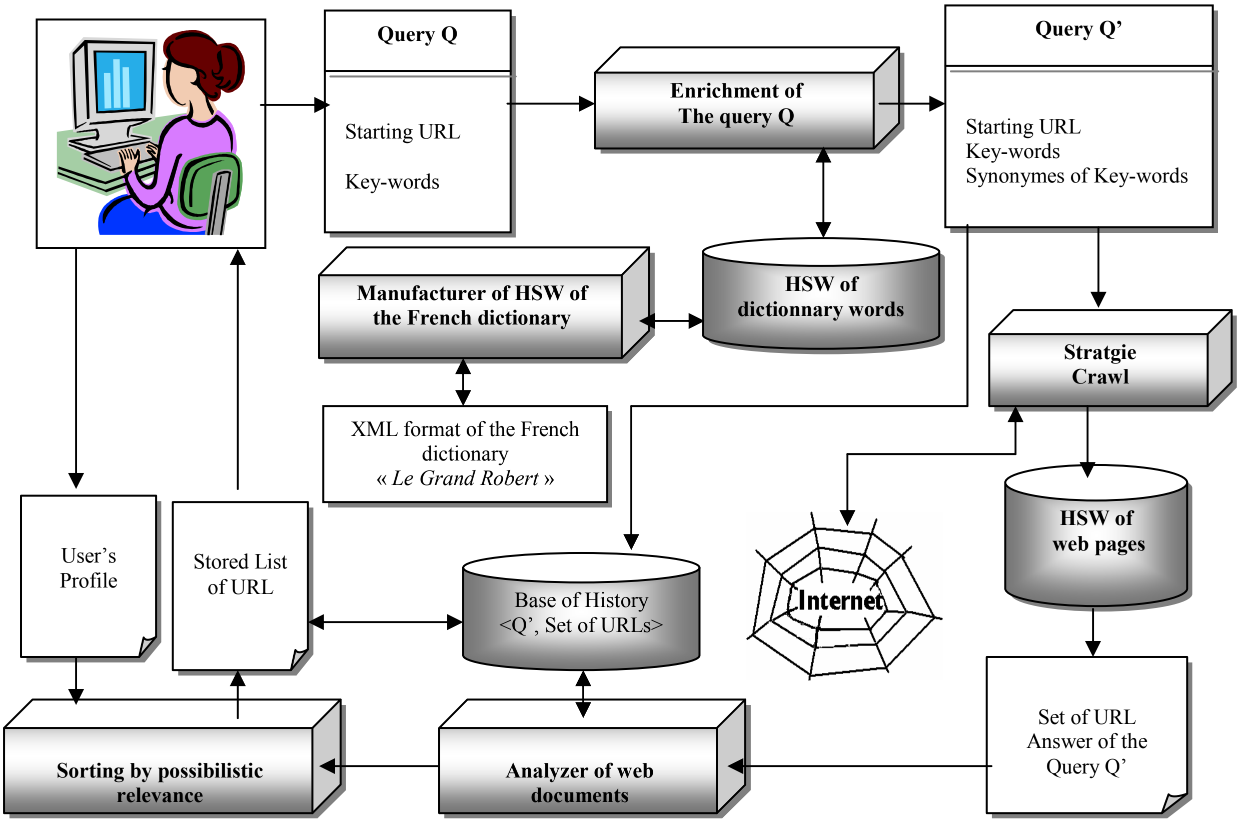
A directed possibilistic network on a variable set V is characterized by a graphical component and a numeric component. The first one is a directed acyclic graph. The graph structure encodes independence relation sets just like Bayesian nets (Borgelt et al., 2000) (Ben Farhat et al., 2002). The second component quantifies distinct links of the graph and consists of the conditional possibility matrix of each node in the context of its parents. These possibility distributions should respect normalisation. For each variable V:

- If V is a root node and dom(V) the domain of V, the prior possibility of V should satisfy: $\max_{v \in dom(V)} \Pi(v) = 1$;
- If V is not a root node, the conditional distribution of V in the context of its parents context should satisfy: $\max_{v \in dom(V)} \Pi(v | Par_V) = 1$; $ParV \in dom(Par_V)$
  where: dom(v): domain of V ; ParV : value of parents of V ; $dom(Par_V)$: domain of parent set of V .

4. GENERIC ARCHITECTURE OF SARIPOD SYSTEM

To support the re-use, we chose the modularity for SARIPOD system. The co-operation of the various modules makes it possible to generate the generic architecture of our system composed of six modules (see figure 1).

Figure 1. Generic architecture of SARIPOD system


Query Q
Starting URL
Key-words
Enrichment of The query Q
Query Q'
Starting URL
Key-words
Synonymes of Key-words
Manufacturer of HSW of the French dictionary
HSW of dictionnary words
XML format of the French dictionary « Le Grand Robert »
Stratgie Crawl
HSW of web pages
Internet
User's Profile
Stored List of URL
Base of History <Q', Set of URLs>
Set of URL Answer of the Query Q'
Sorting by possibilistic relevance
Analyzer of web documents

4.1 HSW dictionary module

Our approach consist in representing the dictionary by a graph (HSW: there is an arc of a top A towards a top B if and only if the entry B appears in the definition of entry A as a synonym). Consequently, the problem of synonymy in a dictionary is brought back to a study on the graphs seeking to exploit the networks thus established between the words. It acts, very often, to detect components having of the specific properties in terms of graph such as click them or the components N-related thus leading to the regrouping of synonymous words.

Our data base is the French dictionary "*Le Grand Robert*" with XML format in which the elements are described by a whole of beacons allowing each one to associate semantics the various components. The enormous volume of corresponding XML file has constrained us to reduce the size of this dictionary by keeping only information which interests us to make the treatment more effective (Abdallah et al., 2003) (Awada et Chebaro, 2003). This was done, obviously, without loss of relevant information. In fact, we limit ourselves only to the nouns, verbs, adjectives and adverbs of this dictionary, which exempt to us blank words such as the pronouns, the definite articles and indefinite, the prepositions, the auxiliaries (to be and to have), etc.

4.2 Enrichment of the user's query module

The enrichment of user's query is a paramount stage in SARIPOD system. Indeed, this module sends user's keywords to HSW dictionary module which determines their synonyms and expends them to the enrichment of the user's query module. With the reception of these synonyms, this module publishes consequently, a new query Q' containing, in addition to the starting URL chooses by the user of system, the set of keywords and their synonyms. In fact, the new query Q' is well enriches by news words.

4.3 Strategie crawl module

We propose within this framework a systematic crawling technique via TBC algorithm, whose scenario is as follows :

1. While a page on two contains the word M it is necessary to continue to visit the outgoing pages of this page (whatever the depth) for that would be nevertheless a shame not to go to visit these outgoing pages which have a very strong probability of containing the word M;
2. When two successive pages do not contain the word M (whatever the depth), we stops research in this branch, so we need backtrack then.

4.4 Web document analysis module

This module allows the extraction of the logical structure web document. We mind to store such a document in an editable and exchangeable format that represents explicitly its structure and its content. The strategy of this module is based on the ticketing method. It is composed of several analysis steps that leads to the transformation of the document in a logical structure where each text block has a level and a label that represents explicitly its logical role (Bounhas et al., 2006). Searching the user's key-words in such a structure of document and not in the whole document proves the qualitative character of the SARIPOD system.

4.5 Sorting documents by possibilistic relevance module

We allotted a coefficient of relevance to each logical structure according to its importance in the web document. These coefficients are calculated with the following way:

$$\alpha_{ML} = ML + \mathrm{Max}(\alpha_{Legends}, \alpha_{P}),$$
$$\alpha_{Li} = ML - L_i + \mathrm{Max}(\alpha_{Legends}, \alpha_{P}),$$

Where ML is the maximal level, L_i is the level *i* of logical structure and P is paragraph.

The quantitative relevance of each logical structure of a web document of the collection, with the query is $Q = (t_1, t_2, \ldots, t_T)$, is calculated in the following way:
The expression of $\Pi(LSd_j|Q)$ is then proportional to:
$\Pi'(LSd_j|Q) = \Pi(t_1| LSd_j)*\ldots* \Pi(t_T| LSd_j) = nft_{1j} *\ldots* nft_{Tj}$
Where $nft_{ij} = tf_{ij} / \max (tf_{kj})$: the normalized frequency of the terms of the query in the logical structure.
The certainty to restore a logical structure (LS) of a relevant document d_j for a query, noted $N(LSdj |Q)$, is given by: $N(LSd_j|Q) = 1- \Pi (\neg LSdj|Q)$
Where: $\Pi(\neg LSd_j|Q) = (\Pi(Q|\neg LSd_j)* \Pi(\neg LSd_j))/\Pi(Q)$
In the same way, $\Pi(\neg LSd_j| Q)$ is then proportional to: $\Pi'(\neg LSd_j|Q) = \Pi(t_1| \neg LSd_j)* \ldots*\Pi(t_T|\neg LSd_j)$
This numerator can be expressed by: $\Pi'(\neg LSd_j|Q) = (1- \phi LS_{1j})*\ldots* (1- \phi LS_{Tj})$
Where: $\phi LS_{ij} = \mathrm{Log}_{10}(nCLS/nLSd_i)*(nft_{ij})$
And :
nCLS = The number of logical structure of the documents of the collection,
$nLSd_i$ = The number of logical structure of the documents of the collection, containing the term t_i,
So: $DRMPLS_j(d_i) = \Pi(LSd_i|Q) + N(LSd_i| Q)$
Let us note the degree of relevance mixed possibilistic (DRMP) of the document d_i by:
$DRMP(d_i) = \sum_j (\alpha_j * DRMPLS_j(d_i))$

The preferred documents are those which have a high value of $DRMP(d_i)$. Let us note that the α_j are parameterized in our system and can be modified according to the user's profile. Indeed, if we seek, for example, to the documents containing the word "M" in figures, it is enough to give the greattest importance to the coefficient of relevance corresponding to the figure legend appears (α_{FL}). Consequently, $DRMP(d_i)$ of these documents will be most significant and will be posted in the heading of the list result of sorted documents.

Example: Assume a three document collection containing the four terms t_1, t_2, t_3 and t_4:
$d_1 = \{t_1, t_1, t_1, t_2, t_2, t_3\}$, $d_2 = \{t_1, t_1, t_2, t_2, t_2, t_2\}$ and $d_3 = \{t_1, t_3, t_3, t_3, t_3, t_4, t_4\}$
These terms are distributed in the logical structures of these three documents as table 2 indicates. The degree of relevance mixed possibilistic of each document d_i is $DRMP(d_i)$.

Table 2. Distribution of the terms in the logical structures of the three documents

logic Structure of document	**d_1**	**d_2**	**d_3**
Maximal Level (ML)	t_1	t_1, t_2	t_4
ML-1	t_2		t_1, t_3
ML-2			t_3
ML-3		t_2	
ML-4	t_3		
Figure Legend (FL)			t_3
Table Legend (TL)	t_2		
Multimedia Sequence Legend (MSL)		t_2	
paragraph (P)	t_1, t_1	t_1, t_2	t_3, t_4
DRMP (d_i)	**11.28**	**11.76**	**19.07**

For the example of this query Q, the document d_3 is more preferred than documents d_2 and d_1. We notice that the most relevant document is that whose query's terms exist in its logical structures having the significant coefficients of relevance α_j, such as the maximum level (e.g., the principal title of the document), Maxlevel -1 (e.g., title 1), Maxlevel -2 (e.g., title 2), etc. Thanks to our mixed possibilistic approach, we also noticed that: even if the selected terms tend to select this document, these terms are not most frequent of the document (t_4 is not the most frequent term of d_3).

4.6 Optimization system module

The optimization system module of SARIPOD allows to its users a significant profit in term of response time of our system. Indeed, this module makes it possible to build a base of history of the querys and their answers, already passed by the system. With the reception of a new query, the system consults this base of history, seeks the nearest query in this base and finally, it up date the answer by eliminating URLs not available on the web and by adding nonexisting news URLs in this base of history. This requires the expend of the new query to the strategie crawl. This new obtained answer will be useful, in the same way, like history for later querys. Let us note that the user will be able to use the result of the history directly or to change his profile while launching a new sorting of the documents according to his new parameters.

5. CONCLUSION

This paper presents an Internet information retrieval system based on Hierarchical Small-Worlds (HSW) and Possibilistic Networks (PN). The first HSW, for the words of the French language, is used to take account of the dependences between these words. The second HSW is devoted to the web pages required and translated in the same way the dependences between these pages. The Possibilistic Networks generates the mixing of these two HSW in order to organize the sershed documents according to the user profile.

This paper presents a mixed approach for Information Retrieval Model using possibilistic logic. The approach proposed by Brini and al. (Brini et al., 2004) is only based on the quantitative setting of possibility logic, our mixed approach extends it by also focusing on the ordinal setting. In fact, we add a qualitative approach which introduces preferences between retrieved documents. A user of our system has the possibility, then, to have his retrieved documents in any wished order via an interface which allows him to choose his own parametres of preferences. This approach had been implemented within Java language and test results were in conformity with our initial needs. As future works, we will firstly show how this model based on possibility logic could be a counterpart to Bayesian probabilistic models. We will also plan to consider preferences between query terms.

REFERENCES

Abdallah, H., Sleiman, R., et Harati, A., 2003. Etude de la synonymie dans les dictionnaires et réalisation d'un outil de mesure de la proximité de sens. *Mémoire de fin d'études de maîtrise d'informatique*. Université libanaise, faculté des sciences I.

Awada, A., et Chebaro, B., 2004. Etude de la synonymie par l'extraction de composantes N-connexes dans les graphes de dictionnaires. *JEL2004*, Nantes, France.

Benferhat, S., Dubois, D., Garcia, L., et Prade, H., 2002. On the transformation between possibilistic logic bases and possibilistic causal networks. *Int. Journal of Approximate Reasoning*, 29, n°2, 135-173.

Borgelt, C., Gebhardt, J., and Kruse, R., 2000. Possibilistic Graphical Models. *Computational Intelligence in Data Mining*. CISM Courses and Lectures 408, 51- 68.

Bounhas, I., Habacha, A., et Ben Ahmed, M., 2006. XSLS :Une approche basée sur l'étiquetage pour la structuration logique des documents scientifiques Web. *Mémoire de Mastère en informatique*, Ecole Nationale des Sciences de l'Informatique, à publier en Tunisie.

Brini, A. H., Boughanem, M., et Dubois, D., 2004. A possibilistic approach to information retrieval. IRIT, Université Paul Sabatier.

Brini, A. H., et Boughanem, M., 2003. Relevance feedback : introduction of partial assessments for query expansion. *Proc. 2d. EUSFLAT Conf*, Zittau, Allemagne, 67-72.

Douglas, W., et Houseman M., 2002. The navigability of strong ties: small worlds, ties strength and network topology. e*Scholarship Repository*, University of California.

Dubois, D., et Prade, H., 1987. *Théorie des Possibilités : Application à la Représentation des Connaissances en Informatique*. Edition MASSON. Paris.

Gaume, B., Hathout, N., et Muller, P., 2004. Désambiguïsation par proximité structurelle. *TALN 2004*, Fès, Marroc.

Watts, D., and Strogatz, S., 1998. Collective dynamics of 'small-world' networks. *Nature*, (393), 440–442.

Zadeh, L. A., 1978. Fuzzy Sets as a basis for a theory of Possibility. *Fuzzy Sets and Systems*. 1 :3-28.