Journal of Theoretical and Applied Information Technology
15 August 2012. Vol. 42 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

NOISE ABATEMENT IN THE ARABIC IRS RESULTS BY APPLYING A SENTENCES MORPHOSEMANTIC FILTER (GENE FILTER METHOD)

[1]Adil ENAANAI, [2]ABDELAZIZ DOUKKALI, [3]BOUBKER REGRAGUI
[1]Ph. D Student., TIES team, ENSIAS, MedV Souissi University, RABAT
[2]Assoc. Prof., TIES team, ENSIAS, MedV Souissi University, RABAT
[3]Assoc Prof., TIES team, ENSIAS, MedV Souissi University, RABAT

E-mail: [1]enaanai_adil@yahoo.fr , [2]doukkali@ensias.ma , [3]regragui@ensias.ma

In this paper, we introduce a new approach to facilitate the calculation of relevance and noise abatement in information research systems in Arabic language. Our method is to remove morphosemantic ambiguity due to agglutination and lack of vocalization of the Arabic words. To do, we have proposed to transform words to semantic gene. The latter represent an accurate determination of the word meaning. They contain the type, context, definition and vocalized shape of all possible cases may be taken in the Arabic word. In our approach we consider all possible meanings of the terms by applying a morphosemantic variation based on a recursive algorithm. Obtained variants are filtering by using of the sentence context, user profile and the Arabic phrase synthesis rules. The result is a semantically coherent text ready to be used by an information search system.

Keywords: *Semantic Gene; Arabic Disambiguation; TALN; Information Research.*

1. INTRODUCTION

The information research is being an essential service in the current web. The information retrieval systems (SRI) improve in a rapid way to cover the needs of users to a large set of dispatched documents anywhere on the Web. And therefore, more relevance in the research results. The request sent by the user is an expression that determines its needs in several forms (written, spoken or shown). A good SRI must offer all possibilities contribute to the expression of the user intent. The query can be written in several languages and respects different syntaxes. The current SRI use automatic processing of natural language (TALN) to remove some kinds of ambiguity. Then use ontologies to find the exact meaning of the words. This type of treatment has yielded good results, but remains limited to certain special cases such as morphosemantic ambiguity in the Arabic language. At the moment, there is no function of similarity able to guessing the exact meaning of some Arab words, because the Arab sentences arise in different contexts or morphological variants of multiple meaning.

Since almost of the Arabic documents on the Internet are unvowlized. Information search systems are required to guess their vocalization to find the good sense of the words. There is researchs that offer to write Arabic queries in Latin letters to clarify the vocalization user, but it mean nothing if the documents on the Internet are not also translated to Latin letters. In addition, this last proposal will bring no added value to the Arabic language as a living language. The overall objective of research in Arabic language is to find the relevant documents to the user intention while rehabilitating the Arabic language value. In this paper, we will define a new method to clarify the meaning of the Arabic sentences to be used in the information research systems. This method is focused around a multi-level analysis based on morphosemantic appearance of the words, the context of phrases and the user profile.

2. PROBLEM

In the information research systems, the relevance is function of degree of similarity between the request and the document. However, several functions of similarities are offered currently. Most of these functions are based on the principle of the vector distance where the meaning of the words is not supported. However, two words whose the distance is zero are similar. However, there are words of high distance which mark the same meaning (synonyms) or words of distance equal to zero, which means several things. Functions based on the vector distance are unable to

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

provide the exact value of the semantic similarity of words. In addition, there are algorithms of rooting which contribute to the calculation of relevance by comparing the roots of the words which gave good results. But they remain insufficient, because there are some Arabic words whose roots are written in the same way but their meanings are different. The difficulty of having a function of semantic similarity lies in the fact that the comparison of the meaning between two words is possible that after you include a valid morphosemantic analysis. That's why we need database which give us more information that can assist in the realization of this type of analysis.

To give more precision to the application of the similarity function, we chose to solve three problems: the morphological ambiguity, the semantic ambiguity and the function of similarity.

The morphological ambiguity: This is the type of ambiguity due to the clumping of words, where the articles, prepositions and pronouns stick to adjectives, nouns, verbs and particles which they relate.

Example1:

The word "بطريق" mean two different concepts:

Concept1: **بطْرِيقِ** (penguin)

Concept2: **بِ طَرِيق** (By a road)

Example2:

The word "ليمون" mean two different concepts:

Concept1: **لِيُمَوّنَ** (To finance)

Concept2: **لَيْمُون** (lemon)

The semantic ambiguity: This is the type of ambiguity due to the unvowelization or the type of word in the phrase, where an unvowlized word can have several variants semantically different.

Example1:

The word "كتلب" mean two different concepts:

Concept1: **كِتابٌ** (Book)

Concept2: **كُتَّابٌ** (Pre-school)

Example2:

The word "عادل" mean two different concepts:

Concept1: **عادل « Just »**

(Object in the sentence «عادل إمام » « Just Imam » an Arabian actor)

Concept2: **عادل « Just »**

(Property in the sentence «إمام عادل » "Just IMAM")

The similarity: The similarity between two texts is the objective of all IRS. However, each language has its own properties. Therefore, a function applied to the French or English is not necessarily efficient for the Arabic language.

Example:

Consider the two sentences:

Sentence1= « ذهب بطريق القطب الجنوبي »

Sentence2= « ذهب بطريق القطب الجنوبي »

***Sim*(**Sentence1**,** Sentence2**)** = 100%

The famous functions of similarity used to compare two sentences give the value 100% which means that Sentence1 and Sentence2 are similar. But on considering the morphosemantic variation of words, the two sentences are not similar, because its can have the following meanings:

Phrase1= « ذَهَبَ بِطَرِيقُ القُطْبِ الجَنُوبِي »

Phrase2= « ذَهَبٌ بِ طَرِيقِ القُطْبِ الجَنُوبِي »

Hence, ***Sim*(**Sentence1**,** Sentence2**)** =**??? <100%**

3. STATE OF THE ART

The calculation of relevance is a process based on complex calculations. According to selected criteria of relevance, we can define several approaches based on a function of relevance. In the remainder of the State of the art, we present the various criteria for the calculation of the relevance and the famous functions of similarity.

A. The relevance criteria

The classification of the results is specific to each engine algorithm, that is, a method based both on logical and mathematical criteria to give a score to a couple page-request. If a motor returns 300 000 results for a query, the list of results is classified from the first to the 300 000th by this method of scoring.

A first sorting is done through playoffs criteria permitting the engine to determine whether a page should be removed or not in the list of results. For example the language, when the engine uses linguistic filters.

The final sorting is the result of a criteria combination which will allow assigning to each page a score to the search query. Each of these criteria is designed to measure the relevance of a page for a query.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Most of the relevance criteria evaluations are related to the content of the page, but certain criteria are linked to the site as a whole [3].

1) Criteria related to the content of the page ("in-page"):

- Content of title
- Frequency of the keywords
- Density index
- Contents of the URL
- Proximity and order of keywords
- Size and styles of fonts
- Presence in the Meta Keywords
- Weight in KB of the page
- Date of creation / modification

2) Criteria relating to the site ("off-page"):

- Domain name
- Popularity
- Theme of site
- Size of the site
- Click index

3) The criteria of trust:

The confidence index, the same title that the index of popularity on the internet are assigned by search engines across hundreds of criteria necessary for the positioning of your website in the results of the engines [3]. Here are the best known:

- The site security
- Duration of registration of your domain name
- Identifiable physical coordinates
- Existence of legal notices
- Date of editing pages
- Frequency of the site updating
- Renewal of new articles or pages
- Geographic location of your host
- The image hosting server response time
- Free code errors (label W3C)
- Step content hidden in the code and not visible by the user
- Free content

4) The popularity criteria

- Architecture of internal links
- Index of popularity of your external links sites
- Using a file of exploration for engines (sitemap)
- Internet traffic, the number of visitors
- Time spent on the page by users
- Number of page views by visitor
- Use of the html/xhtml code mostly

B. The function of similarity

1) Damerau–Livenshtein algorithm

In theoretical computer science and computer science, the Damerau-Levenshtein distance is a distance between two strings. We calculate the minimum number of operations necessary to convert a string to another, where a transaction is defined as the insertion, deletion, or substitution of a single character, or as a transposition of two characters. Frederick j. Damerau has not only distinguished these four operations of Edition, but also stated that they correspond to more than 80% of human misspellings. The Edit distance was introduced by Vladimir Levenshtein, who then generalized this concept with multiple operations of Edition, but without including rearrangements.

Example

If M = « محمود » and P = « محمود »

Then LD (M, P) = 0, because no operation was made

if M = « محمود » et P = « محدود »

Then LD (M, P) = 1, because there was a remplacement (change of 'م' to 'د').

2) The Jaro-Winkler method

Jaro-Winkler distance measure the similarity between two strings. It is a variant proposed in 1999 by William e. Winkler, from the distance of Jaro (1989, Matthew a. Jaro) which is mainly used in the detection of duplicates.

More the Jaro-Winkler distance between two strings is higher, the more they are similar. This measure is particularly adapted to the treatment of short chains such as names or passwords. The result is normalized to have a measurement between 0 and 1, zero representing the absence of similarity.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Consider two strings المرجع and المعرج. The correspondence table is:

	ا	ل	م	ر	ج	ع
ا	**1**	0	0	0	0	0
ل	0	**1**	0	0	0	0
م	0	0	**1**	0	0	0
ع	0	0	0	0	**1**	0
ر	0	0	0	**1**	0	0
ج	0	0	0	0	0	**1**

The Jaro distance is : $d_j = \frac{1}{3}(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{m})$

$$\boldsymbol{d_j = \frac{1}{3}\left(\frac{6}{6} + \frac{6}{6} + \frac{6-1}{6}\right) = 0,944}$$

Where:

$|S_i|$: is the string length

m: is the number of matched characters

t: is the number of transpositions

3) The Jaccard method

The Jaccard index (or Jaccard coefficient) is the ratio between the cardinality (size) of the intersection of the sets considered and the cardinality of the union of the sets. It allows assessing the similarity between the sets.

$$Sim(D,Q) = \frac{\sum_i (a_i * b_i)}{\sum_i a_i^{\,2} + \sum_i b_i^{\,2} - \sum_i (a_i * b_i)}$$

D={a0, a1, a2,…, an} ; Q={b0, b1, b2,…, bn}

Example:

Consider two words:W1=«الكتاب » and W2=«المكتبة».

E= W1UW2={ا,ل,م,ك,ت,ب,ة}

W1={1,1,0,1,1,1,0}

W2={1,1,1,1,1,1,1}

$$Sim(W1,W2) = \frac{5}{5+7-5} = 0,714$$

4) The TF-IDF method

The TF - IDF (of English Term Frequency-Inverse Document Frequency) is a method of weighing often used in research information, especially in the search of texts. This statistical measure helps us assess the importance of a word in a document, relatively to a collection or corpus. Weight increases the number of occurrences of the word in the document. It varies also according to the frequency of the word in the corpus. Variants of the original formula are often used in search engines to assess the relevance of a document based on the user's search criteria.

The theoretical justification for this weighing scheme is based on the empirical observation of the frequency of words in a text which is given by the Zipf law. If a query contains the term T, a document is more likely to respond that it contains this term: the frequency of the term in the document (TF) is great. However, if the term T is itself very common in corpus, that is present in many documents (e.g. articles defined), it is in fact not discriminating. This is why the scheme proposes to increase the relevance of a term with its rarity in the corpus (high frequency of the term in the IDF corpus). Thus, the presence of a rare term of the request in the content of a document increases the "score" of the latter.

4. CONTRIBUTION

The calculation of relevance in our approach is focused on semantic similarity function which gives a result as a percentage of equivalence between two Arabic words. Knowing that they are written in various derived forms, it had to begin by morphological analysis which returns the origin of the derivative in question. Therefore, the possibility of separate affixes of the word is subsequently obtained by the original unvocalized of the word which may refer to several meanings. The probable meaning to be just, is that which is on conflict with the user profile. To filter the true meaning, we have developed an automatic profiling system that brings together user queries and implements format indexed in a database. Our approach has given a good result on the morphosemantic ambiguity. In the remainder of this part of paper we will present the various stages of analysis that we introduced in the relevance calculation [4].

1) The function morphosemantic of similarity (FMSS)

The function of morphosemantic similarity (FMSS) is a function that considers the morphosemantic variations of the word by giving a probability of the morphosemantic similarity (PMSS). This probability is calculated by the principle of Jaccard using after removing all ambiguities in the sentence. We begin by a morphosemantic derivation of the various possible cases of ambiguous words. Then we apply a filter based on the contextual database of Arabic words ARRAMOUZ ALWASEET [1]. This filter returns a contextually consistent sentence.

Journal of Theoretical and Applied Information Technology
15 August 2012. Vol. 42 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Example 1: (The morphologic derivation)

We consider the two words: M1="بحر" and M2= "بحور"

Step 1: We start with a morphological derivation based on the recursive algorithm [2] for all possible cases of morphological meaning. We obtain:

The morphologic variants of M1:

VM(M1) = {**بحر ;ب حر**}

The morphologic variants of M2:

VM(M2) = {**بحور ;ب حور** }

Step 2: For each morphological variant, we apply semantic derivation based on the recursive algorithm using the ARRAMOUZ AL WASEET database. We obtain:

VMS(M1)= {**بَحْرٌ ; بِـ حَرٌ ;بِـ حُرٌ**}

VMS(M2)= {**بُحُورٌ ; بِـ حَورِ ;بِـ حُورِ**}

Note: the word "بَحْرٌ" is repeated twice because it represents two different meanings, the first means "Sea" and the second means "Model of Arabic poetry".

Step 3: We transform the elements of variations to singular form. We obtain:

VMS(M1)= { **بَحْر ;بِـ حَر ;بِـ حُر** }

VMS(M2)= {**بَحْر ;بِـ حَور ;بِـ حُور** }

Step 4: We apply the following formula:

$$\text{PMSS} = \frac{VMS(M_1) \cap VMS(M_2)}{VMS(M_1) \cup VMS(M_2)}$$

We obtain: $\text{PMSS} = \frac{2}{6} = 0{,}33$

This value is low, because we do not know what the context of these two words is. The following paragraph shows how to use the context to refine the results of the morphosemantic similarity function.

2) The context use in the refining of the of similarity's function results

The context of word in the sentence provides more opportunities in the overdraft of the exact meaning. We used the definition database ARRAMOUZ ALWASEET to extract the definitions of words, and finally to deduct the context. In all of the variants, we note that there are words whose context is different the text context. Therefore, it must eliminate the inconsistent variations before applying the function of morphosemantic similarity [2].

The word context comes from its definition by using a recursive algorithm. This algorithm iterates through the definition of the word and of the unambiguous words component the first definition in depth.

Example:

Consider Word M = "مطرقة"

Its definition is "هي آلة حرفية يستعملها النجار لدق المسامير في الخشب"

Therefore, the context of level 1 of the Word 'M' is:

مطرقة
آلة
حرفة
نجار
دق
مسمار
خشب

Consider the definition of each term of the level 1: From these definitions, we construct the context of level 2

آلة	أداة- صناعة - شيء
حرفة	عمل- إنسان- دخل
دق	حركة- اصطدام- جسم- تأثير
مسمار	أداة- نجار- تثبيت- خشب
خشب	مادة- شجرة- كرسي- مائدة- دولاب

The general context is the union of the above contexts:

مطرقة	آلة-حرفة-نجار-دق-مسمار-خشب-أداة-صناعة-شيء- عمل- إنسان- دخل- حركة- اصطدام- جسم- تأثير- تثبيت- خشب- مادة- شجرة- كرسي- مائدة- دولاب

The context is a set of words constituting the namespace containing the concerned Word. The method of disposal of inconsistent variations is done according to the following formulation:

Consider $M_1 \in T_1$

With CT_1: The contexte of the texte T_1

And CM_{1i} : The contexte of the i th variant of M_1

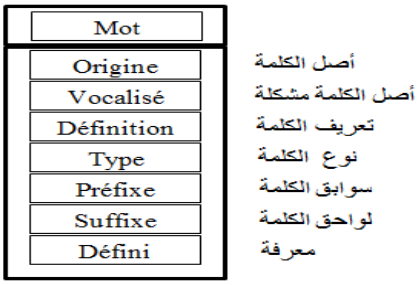
If $CM_{1i} \cap CT_1 < k$ Then delete M_{1i} variant

3) The approach application in a meta-search engine

Journal of Theoretical and Applied Information Technology
15 August 2012. Vol. 42 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

To test the effectiveness of our work, we have developed a meta-search engine. The approach is applied to the documents and the query. The document is transformed into semantic genes containing all relevant information to infer the meaning of a Word [2][6].



A) The document refining

The document refining is a process that transforms the ambiguous words on semantic genes, and then eliminates the inconsistent morphosémantic variants.

Example of context refining of the document:

Sentence = "استخدام البحور الشعرية"

Semantic gene transforming and context refining:

Sentence= شعر بَحر استخدام

B) The query's refining

It is the process which eliminates the inconsistent variants based on the user profile. The profile is a set of invariant words sent by the user. Therefore, they construct a context summarizing its interests. This context is using in the same principle of the documents refining. The method used to define the user profile is presented in the following paragraph [6].

4) The user profiling

There are a countless opportunities offers Internet to private companies, boards advertising and search in terms of profiling of Internet users, for commercial engines. In the vast majority of cases, this trace remains anonymous. And having such a method of profiling is a basic brick to better recognize customers. There are two types of profiling:

Manual profiling: where the user of service must complete a form by answering some questions for the service of property interests of the user.

Automatic profiling: where the user is not invited to complete a form. These are the sent queries that make up the user profile. In this case, the profile can be initialized each time by deleting cookies, because they retain traces of the user. Thus, by removing cookies, it also removes the history of requests [1] [12].

To test the automatic profiling, we developed a meta-search engine for automatic profiling. The concept is to create a profile for each user. Profile is used to infer a user interests based on the requests sent by the latter. The server receives the request from the user. Then, we test if the user has the cookie file created by the server or no. If the user has the cookie file, the server takes the ID of the user from the cookie, and seeks the user's profile in the database. Otherwise, we create a new file cookie for the user. If the server finds that the client identifier exists, it explores its profile built from the semantic entities (SE), and adds the newest sent SE to the set of ES constituting the profile. Then we call the similarity function which calculates the degree of similarity between the document, request and profile. The following figure illustrates the principle of automatic profiling [1].

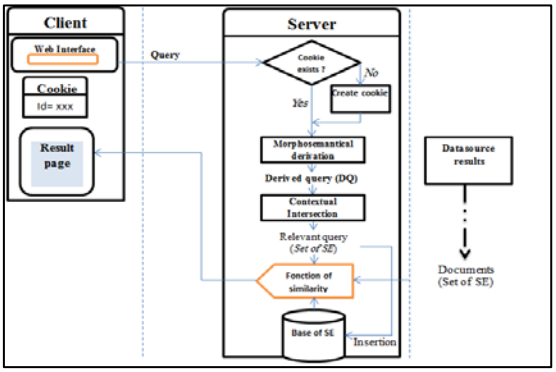


Figure 1: Diagram of the different stages of automatic profiling

5) Two words queries Analysis

The expression of need is made by a multi-word query. The information system research returns a set of documents that contain all of the semantically valid sentences. These sentences include those that contain the desired word. This Word can mean several things, Hence the problem of semantic ambiguity. To reduce the effect of this type of ambiguity, we designed a semantic filtering system that recognizes the type of the word based on the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

rules of constitution of the Arabic sentences. Given the difficulty of semantic analysis of the Arabic sentences, we consider the case of significant sentences of two lemmas.

The two words sentences meet the canonical form of the Arabic writing sentences. Where, we have designed a set of patterns covering almost all of the forms. These patterns are considered as the mussels. The following table shows the different cases of a sentence of two words semantically consistent [9].

Table 1: Types of words

Object		Property		Fact	
Type	Acronyms	*Type*	Acronyms	*Type*	Acronyms
صيغة مبالغة	O_{Mob}	اسم تفضيل	P_{Taf}	مصدر	F
منسوب	O_{Man}	اسم مفعول	P_{Maf}		
جامد	O_{Jam}	اسم فاعل	P_{Fa}		
اسم مفعول	O_{Maf}	صفة مشبهة	P_{SM}		
اسم فاعل	O_{Fa}	صفة	P_S		
		منسوب	P_M		

Note: Just the types listed in the table above are considered.

Table 2: Table of the synthetic mussels

Prefix of the word1	Word1	R	Word2	Example	Pattern Mot1 R Mot2	Relation between Word1 and word2
(ك، ل، و، ب)	D	∅	D	المدينة القديمة	O P	M2 is P for M1
				الإعادة البطيئة	F P	M2 is P for M1
(ك، ل، و، ب)	I	∅	D	مدينة العرفان	O O	M2 is a specification for M1
				طرح التساؤلات	F F	M1 is an action Applied on M2
				تسيير الشركة	F O	M1 is an action Applied on M2
				أرقى العائلات	P O	M1 is P for M2
				أضعف الاحتمالات	P F	M1 is P for M2
(و)	D	∅	I	المسألة صعبة	O P	M2 is P for M1
				التسارع بطيء	F P	M2 is P for M1
(ك، ل، و، ب)	I	∅	I	سنة سعيدة	O P	M2 is P for M1
				تفكير سليم	F P	M2 is P for M1
				تزويج قاصر	F O	M1 is an action Applied on M2
				أقوى رجل	P O	M1 is P for M2
				أصعب اختبار	P F	M1 is P for M2
(ك، ل، و، ب)				شيماء و فاطمة	O و O	M1 in context with M2

We note that there are prohibited cases same as (" O و P"). Therefore, we have designed a set of mussels forming all possible cases of the sentences of two lemmas. This set of mussels is an array of objects where each element describes a phrase (pattern) model. The process of correction is applied firstly on the list of the semantic entities (alimented query) of the user to remove the inconsistent morphosemantic variants [14]. Then, we send the remaining lists for contextual correction system. The latter uses the contextual corpus to refilter the list. The result is one or more lists of consistent semantic entities at the contextual level as at the semantic level. Finally, the research system is receives a suite of semantic genes containing all information that can help the extraction, selection and filtering of relevant documents [10].

There are words that can be objects or properties. Our approach supports the gene by assigning the type of the word.

Example:

Let's say that we have the sentence

ph = "الحاكم العادل" (En: Just governor).

Word1= « الحاكم »; Word2= « العادل »; R= « ∅ »

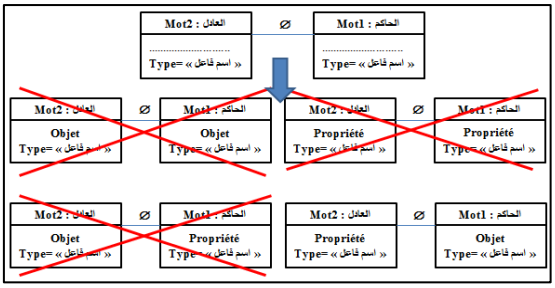


Figure 2: Two words query analysis

The table of mussels shows that the only case which exists in is: Word1 = "object"; Word2="property". Therefore, the sentence " العادل الحاكم" is semantically different to "العادل الحاكم", because its components are not similar. In this way, our system will be able to considerate polysemy [11].

This work is an aspect that has largely been addressed to the Latin language (English, French, ...) and even in some work for the Arabic language. Indeed, research based on the user profile to reduce noise and silence in the information research has yielded satisfactory results especially with the modeling of the user profile and the research domain with the notion of ontology. However, the ambiguity in the terms of query cannot guess the domain to choose from. Hence, we must prepare the query to reduce morphosemantic ambiguity, then guess the context from the user profile and create genes to clarify the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

semantic field and the context intended by the user. The following diagram illustrates the various steps of our approach [18].

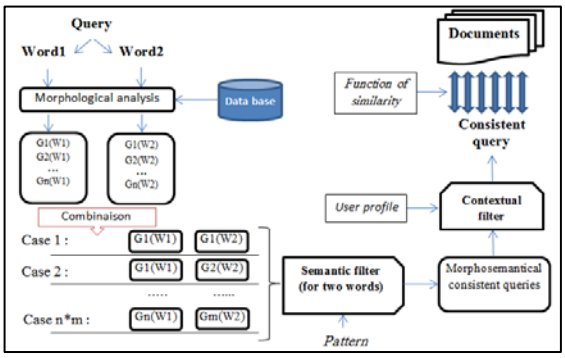


Figure 3: General diagram of approach steps

5. EVALUATION

To test the effectiveness of our method, we have developed a test meta-search engine. The latter uses data sources "Bing", "Yahoo", "Yandex". Then we compared our results with Google results. We have obtained the following table after throwing 100 queries.

Table 3: Evaluation of the approach

	Google	Our system
Average of number of relevent liks sorted in the midst of the ten first positions	**6.47**	**8.14**

6. CONCLUSION ET PERSPECTIVES

In this article, we introduced the concept of the semantic gene that contributes to the Elimination of ambiguity in the information research systems. We also explained how to create the semantic genes from the morphological, contextual and semantic analysis and how to differentiate between homonyms. The automatic profiling is also an interesting factor to approach to the needs of users.

Our target is to automatically create semantic graphs whose semantic genes nodes are very rich in side informational data. Where each node has a context, a definition, a type of Word, a morphological form, a list of successors and a list of predecessors. Finally we wish to develop a meta-research engine which can return optimal results.

REFRENCES:

[1] Adil ENAANAI (2012), An hybrid approach to calculate remevance in the meta-search engines. IJSAT, Vol3 N°2 March 2012.

[2] Adil ENAANAI , A morphosemantic preparation of the Arabic query to improve the calculation of relevance in the IRS, IJCSNS, Vol 12 N°4. April 2012.

[3] BOUGHANEM M (2000), Optimisation de la pertinence dans un SRI : un problème multi-modal approché sous l'angle de la génétique.

[4] RAZAN TAHER (2004), Recherche d'Information Collaborative, Communication de congrès (Toulouse-France), Vol 2.

[5] BAZIZ M (2003). Désambiguïsation et Expansion de Requêtes dans un SRI, Etude de l'apport des liens sémantiques, Revue des Sciences et Technologies de l'Information (RSTI) série ISI, Vol. 8, N. 4/2003, p. 113-136.

[6] BENLAHMAR H. (2006). A New Solution for Data Extraction: GENE/LONE Method, IJCSNS International Journal of Computer Science and Network Security, Vol 6, N° 7.

[7] BEESLEY Ken (1998), Arabic morphological analysis on the Internet. In Proceesing of the 5th International Conference and Exhibition on Multi-lingual Computing, Cambridge, April.

[8] El YOUNOUSSI Y (2011), La racinisation de la langue arabe par lesautomates à états finis (AEF), 4th International Conference on Arabic Language Processing.

[9] ALRAHABI M. (2004). Filtrage sémantique de textes en arabe en vue d'un prototype de résumé automatique, JEP-TALN 2004, Traitement Automatique de l'Arabe Fes.

[10] COLLET K (2003) Méthode du TALN, traitement automatise du langage naturel, notion de l'indexation automatique, Cours, URFIST Bretagne Loire-Atlantique.

[11] ZOUAGHI A. (2004). Une structure sémantique pour l'interprétation des énoncés en langue arabe, JEP-TALN.

[12] KESSLER R. (2008). E-Gen: Profilage automatique de candidatures, TALN2008.

[13] H.DAHMANI (2004) Conception d'un système pour La reconnaissance de mots

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

enchainés arabes, JEP-TALN 2004, Traitement Automatique de l'Arabe, Fès, 20 avril.

[14] P. RESNIK. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. J. Artif. Intell. Res. (JAIR), 11:95–130, 1999.

[15] M. ALJLAYL and O. Frieder, On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach, In 11th International Conference on Information and Knowledge Management (CIKM), November 2002, Virginia (USA), pp.340-347.

[16] A. CHEN and F. Gey : Building an Arabic Stemmer for Information Retrieval. Proceedings of the Eleventh Text REtrieval Conference (TREC 2002). National Institute of Standards and Technology, Nov 18-22, 2002, pp631-640.

[17] K. DARWISH: Building a Shallow Arabic Morphological Analyzer in One Day. Proceedings of the workshop on Computational Approaches to Semitic Languages in the 40th Annual Meeting of the Association for Computational Linguistics (ACL-02), Philadelphia, PA, USA. pp. 47-54.

[18] J.P. DESCLES: Résumé automatique par filtrage sémantique d'informations dans des textes, Présentation de la plate-forme FilText, Technique et Science Informatiques, 2001, nº3, pp. 369-374.