Toward File Consolidation by Document Categorization

Abdel Belaïd and André Alusse

LORIA, Campus Scientifique, B.P. 236, Vandoeuvre-Lès-Nancy, France
{abelaid, alusse}@loria.fr

Abstract. An efficient adaptive document classification and categorization approach is proposed for personal file creation corresponding to user's specific needs and profile. This kind of approach is needed because the search engines are often too general to offer a precise answer to the user request. As we cannot act directly on the search engines methodology, we propose to rather act on the documents retrieved by classifying and ranking them properly. A classifier combination approach is considered. These classifiers are chosen very complementary in order to treat all the query aspects and to present to the user at the end a readable and comprehensible result. The application performed corresponds to the law articles stemmed from the European Union data base. The law texts are always entangled with cross-references and accompanied by some updating files (for application dates, for new terms and formulations). Our approach found here a real application offering to the specialist (jurist, lawyer, etc.) a synthetic vision of the law related to the topic requested.

1 Introduction

With the exponential growth of information available on the Web, the information retrieval becomes increasingly difficult. The search engines have more and more difficulty to satisfy the user's requirements. In response to a request, the document retrieval engines turn over sets built more or less well and ordered according to their criteria of relevance. The experiment showed that neither the relevance nor the linearity of presentation are sufficient factors for the user because 1) they do not make it possible to have a global and synthetic vision result, 2) certain documents can escape the criteria of relevance. Moreover, it is not obvious to write a synthesis in order to constitute a file and it is difficult to follow the cross referenced between documents. This implies to develop useful and efficient tools to assist users in searching documents corresponding to his needs in terms of consultation and organization.

The Project PAPLOO positions in this area. It aims is the definition of a generic framework of transformation and document retrieval for personalized use (document synthesis, folder organization according to the topics, document ranking facilitating their search corresponding to the importance and the quality). The main goal of this project is to help the lawyers of the European Community to synthesize or summarize specific subjects treated in various publication of the European Community (decrees, treaties, rules,…).

H. Bunke and A.L. Spitz (Eds.): DAS 2006, LNCS 3872, pp. 437 – 448, 2006.

A good example being a customs officer intercepting animals transport at the EEC border. He must be able to know the last legislation in use on the animal importation. But the project have other objectives. Avocado, judge, etc. need to build their own consolidated documents, updated automatically, with more or less strong strategy of the appropriation. This implies at least three questions: law classification, crossed consolidation, new text and references.

This paper is organized as follow: in section 2, a brief overview of the whole system is given. Section 3 will be dedicated to file constitution. Experiments and discussions will be done in section 4. Finally, conclusion and future work will be given in section 5.

2 System Overview

The Chain PAPLOO is composed of two distinct parts as shown in Fig. 1. The first part relates to data preparation in terms of OCR (for document images), structure recognition and annotation. The second part concentrates on the constitution of files in terms of reformulation and reformatting. The main language used throughout the chain is XML. Effective research starts after the database constitution enriched by indices (metadata). The user query conditions all the chain. He initialises the total document research (classification, enrichment and reorganization) and allows the constitution of personalized files (documents) by providing the suitable elements of selection. In addition to the personal request, the influence of the user is always present in all the phases of the system through his profile.

The used document are law articles of all kinds belonging to Official Journals (OJ) of the European Union. Documents can be in text format or in PDF which need to be retro-converted. PDF's documents are structured using OCR and retro-conversion processes (Rangoni [1]). In this paper, we will limit our focus on the part related to file constitution.

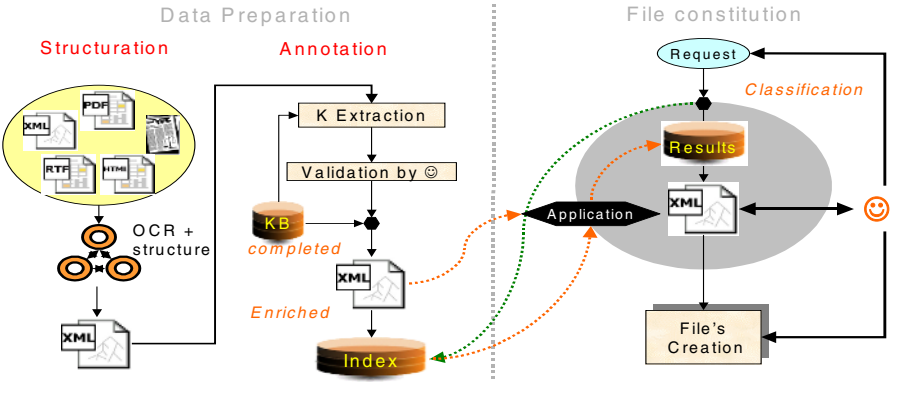

Fig. 1. Chain PAPLOO overview

3 File Constitution

In order to discover sets of similar documents and highlight categories, the categorization, automatic clustering and summarization of documents are possible issues to help the user to solve these problems. A thematic classification allows an organizational vision of the results (Hearst [2]). Moreover, the combination of classification has the potential to provide an accurate, intuitive and comprehensive classified results. Existing work on combining heterogeneous classifiers for information retrieval is widely varied in measures, goals and tasks. Generalization of classifier combination methods were suggested by Lam and Lai [3] and Bennett et. al. [4].

Based on these works, we propose a more dedicated approach for the judicial domain.

3.1 Proposed Solution

The system combines automatic clustering and categorization approaches. Clustering is the process of grouping documents based on similarity of words, or the concepts in the documents as interpreted by an analytical engine. Categorization is the process of associating a document with one or more subject categories.

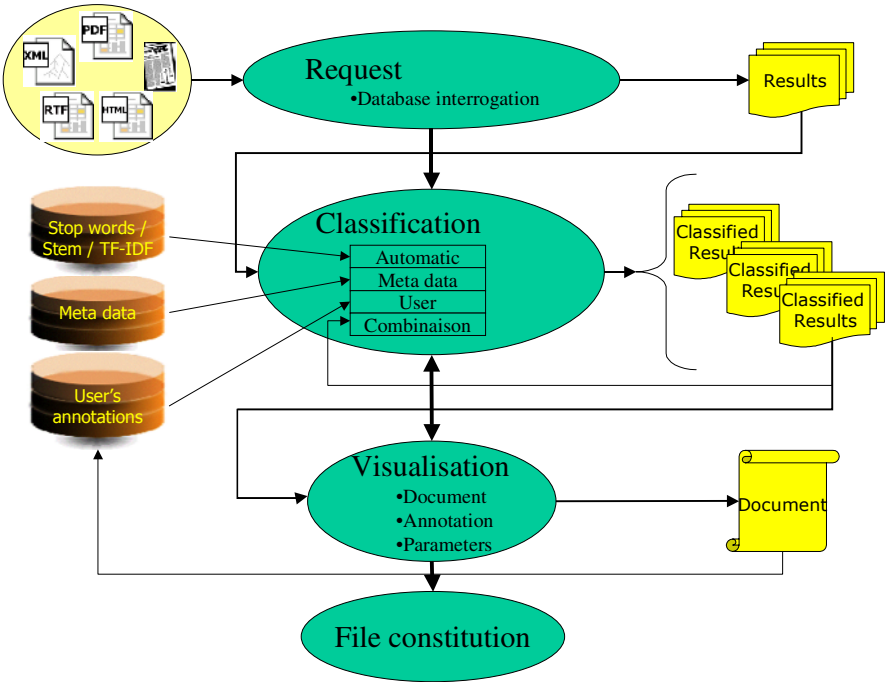

Fig. 2. System overview

In order to discover different point of view on the result set, the system performs several organizations: 1) automatic clustering, 2) metadata categorization, 3) user appreciations. Then, these different classifications are combined: 1) to take into account different points of view, 2) to highlight the main topics (present in the classifiers), 3) to reduce the individual errors, 4) to bring closer the results to the user's concerns.

Fig. 2 shows the system's architecture. First, the user send a request to the database documents Second, he interacts with the different classifications to reorganize the result at glance by acting on the different classifier parameters (removing stop words, performing stemming, pertinence, weight,…). Third, he reads and comments the document by apposing some annotations (finest and personal keywords, more comprehensible by himself…) and at least finalize his file.

Request execution

The documentation base is requested by using a traditional search engine. The query is a set of keywords. The result R is a set of documents where each document d_i is a 4-uplet describing a document result, d_i= (identifier, url, title, summary).

Classification elaboration

Four classifications approaches have been proposed: automatic, using meta-data, using user's annotations and a combination of these three previous classifications.

Automatic clustering: Done with the analysis of the document summaries, three non-supervised statistics-based algorithms have been investigated:

1. Agglomerative Hierarchical Clustering (AHC), Voorhees [5]. Based on the similarity, this algorithm classifies the documents in a hierarchy of classes build two by two. Document/classes similarity is typically computed using a metric such as the Cosine, Dice, or the Jaccard formula. Using a cutting threshold, the AHC produces a C_A classification. The label of classes are build with the most common words shared by the set of documents.
2. Suffix Tree Clustering (STC), Zamir [6]. This algorithm analyses the sentences shared by the documents. Then, it creates a hierarchy. Each sentence constitutes a basic cluster. The sentences are balanced. The score of a sentence depends on the number of words which it contains as well as the number of documents in which it appears. The following stage thus consists in amalgamating these basic clusters according to a similarity function. The same document can belong to several classes. The STC produces a C_S classification.
3. Lingo (Osinski, [7]). Based on the most occurring sentences, it assigns to each one of them the corresponding documents. Each sentence becomes a key label of a cluster. A cluster is validated only if some conditions are satisfied such as: the key word represents a complete sentence, the cluster contains a minimum number of documents, without overlapping, etc. The Lingo produces a C_L classification.

Categorization by metadata: The publication office uses its own thesaurus (or table of index) referring more to the legal topics of this specific base (EuroVoc). The metadata are classified by an expert in the form of a tree of hierarchical indexes. At each leave (i.e. at each path of metadata) will correspond a document classes. As a result of this categorization, a class set C_m is obtained.

User categorization: During document consultation, the user suggests some keywords and scores describing the documents and their interest for his application. For each document d_i, a pair of keywords and associated notes (m_j, n_j) are given. A cross matrix (vector space model, or "bag of word") is established giving for each keyword the associated documents. From this matrix, a similarity or confused matrix (similar to the AHC tree) is computed using Cosine distance on the keywords weighted by the notes. By choosing another cutting parameter, the algorithm produces a C_u classification.

As the user can belong to a group and to a community where the points of view can be completed or generalized, we enlarge the user categorization to the group and to the community. For the group and the community, we put together all the user points of view. This implies to consider all the keywords and to calculate for each one of them the average of the notes given by all the users. The clustering is then similar to the individual grouping. This approach allows the user to share and confront his opinion and the system to have a more global view. This produces two additional classifications: C_G and C_C.

Hence, seven classifications have been obtained by the previous approaches: C_m (metadata), C_A (AHC), C_S (STC), C_L (Lingo), C_U (User), C_G (Group) and C_C (Community).

Classifier combination

These classifications considered individually are not fully satisfying. Each of them is very specialized. To reduce their specific drawbacks, we decided to combine them.

The combination is based on the AHC algorithm with Cosine distance.

Let be C_i a classification, C_i is a set of classes $C_i=\{c_1,c_2,\ldots, c_n\}$ where $c_j=(\omega,\{d_1,\ldots,d_m\})$; ω is the class label and d_i a class document. For the combination, several steps are followed:

All the topics or labels ω are extracted from the all the classifications.

These labels are then used to build the cross matrix M. The label weight depends on its weight classifier and to its occurrence in different classifiers. Weights are assigned to the classifiers according to their accuracy deduced from the experiment. Of course, these values could be refined by the user.

		Documents identifier					
		d1	d2	d3	d4	d5	d6
labels	slaughter	0.0	1.0	1.0	0.0	0.0	1.0
	animal killing	0.0	3.0	0.0	0.0	0.0	1.0
	butchery	0.0	1.0	0.0	0.0	0.0	0.0
	Agricultural activity	1.0	1.0	1.0	0.0	0.0	1.0
	agribusiness	2.0	1.0	2.0	1.0	1.0	3.0
	farm-produce	1.0	1.0	1.0	1.0	1.0	0.0
	aid	3.0	0.0	0.0	0.0	0.0	0.0

Fig. 3. Vector Space Model

Considering this fact, $M_{i,j} = \sum_{k=1}^{n} (d_i \in C_k(\omega_j)) \times p_k$ where p_k is the weight assigned to the classification algorithm and $d_i \in C_k(\omega_j)$ is a binary function (the class ω_j contains or not the document d_i). Fig. 3 shows an example of a cross matrix (vector space model) where the columns represent the 6 document (d_i), while the lines represent the labels (slaughter, animal killing …). The vector representation for the document d_1 is: V_{d1} = (0.0, 0.0, 0.0, 1.0, 2.0, 1.0, 3.0).

Then, the similarity matrix is computed using the Cosine method (1):

$$D(X,Y) = \frac{\sum_{i=1}^{t} x_i y_i}{\sqrt{\sum_{i=1}^{t} x_i^2 \cdot \sum_{i=1}^{t} y_i^2}} \quad (1)$$

Finally, we build the hierarchical tree, based on dendogram, a binary tree structure, with the leaves being the individual document points, the internal nodes being (partial) clusters, and the arcs recording the distance between any two (partial) clusters or documents.

Fig. 4 shows the similarity matrix before and after the classification.

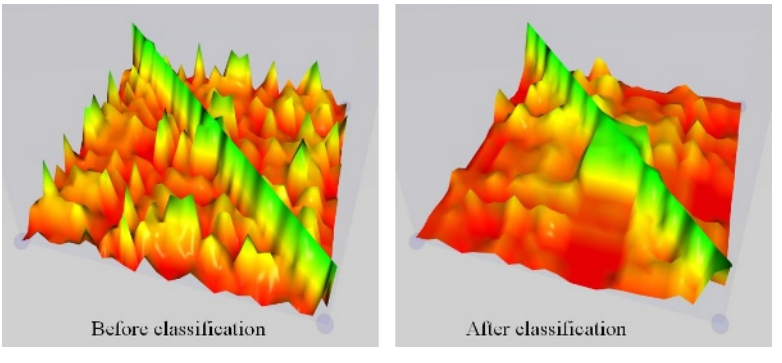

Fig. 4. Similarity matrix

3.2 Classifier Evaluation

In order to help the user to build his file and to adapt the classifier's parameters to solve his needs in a fair way, two objective measurements revealing the cluster qualities were defined. Based on Lamirel [8], concepts of precision and recall used in documentary engineering were adapted to the evaluation of classifications of the documents. The precision measures the percentage of relevant documents among those turn over by the query (N is the number of results, P is the number of relevant documents then Precision = P/N). The recall measures which is the proportion of relevant documents turned over by the system (P is the number of relevant documents retrieved by the system, R is the number of relevant documents in the database, then

Recall = P/R). For our system, the precision measures the classes homogeneity and the recall measures the classes independence.

These table (Fig. 5) shows the classification organisation composed by set of classes. Each class contains documents. Words describe the documents.

		Words				
Class	Documents	t_1	t_2	t_3	…	t_n
C_1	d_1		1			1
	d_2	1	1			1
C_2	d_3			1		
	d_4	1	1	1		
C_n	d_{n-1}	1				1
	d_n			1		

Fig. 5. Repartition of words in documents and in classes

First, the word precision (t_i) in the class C_j is performed as follow (2):

$$P_{C1}(t_1) = \frac{nbDoc(t_1 \in d)}{nbDoc(C_1)} \tag{2}$$

(= 1 when all documents of class C contains t).

Then the class precision is determined (3):

$$P_{C_1} = \frac{\sum_{1}^{n} P_{C1}(t_i)}{|C_1|} \tag{3}$$

(=1 when documents contains the same terms, class is homogeneous, documents are described with same terms).

Finally, the classification precision is determined (4):

$$P = \frac{\sum_{1}^{n} P_{Ci}}{|C|} \tag{4}$$

(=1 means that the division is done in homogeneous classes).

Similarly, the word recall in class is determined (5):

$$R_{C1}(t_1) = \frac{nbDoc(t_1 \in d, d \in C_1)}{nbDoc(t_1 \in d, d \in \{C_i\})} \tag{5}$$

(=1 means that the word t_i appearing only on documents of class C_j).

The class recall is performed as follow (6):

$$R_{C1} = \frac{\sum_{1}^{n} R_{C1}(t_i)}{|C1|} . \quad (6)$$

(=1 means that none of the terms constituting C_j belongs to the other classes).
Finally, the classification recall is determined (7):

$$R = \frac{\sum_{1}^{n} R_{Ci}}{|C|} \quad (7)$$

(=1 means that the division is done in independent classes).

4 Experiments and Discussion

The method is experimented on a small part of documents of the EC (2000 documents including 453 rules, 368 written questions, 242 treaties, …) which are enough representative to validate the classification approaches.

Fig. 6 shows an example of automatic classification produced by the « lingo » classifier. The request given by the keywords "animals" turn over 120 documents. 17 classes were detected, (the class named "publication of request" contains 14 documents). The sentences which are use to define the name of classes are shared by the documents. Some documents belong to more than one class. In this case, 21 documents are duplicated (141 – 120) . 21 documents are unclassified (others). Fig. 6 shows also an example of AHC classifier and an example of classification combination. The user can discover different organisation of the result sets. The system can help him to explore the documents while proposing different approaches and different topics.

Fig. 6. Example of classifications

Then, in order to enlarge the algorithm comparison, three queries with keywords: “animals”, “rice’ and “animals transport” are requested. For each classifier, we give the number of documents return by the search engines (Res), the number of classes detected (Cl), the total number of documents classified, one document could be member of several classes (Nb), the number of document unclassified (Un), the assignment coverage Co and the overlap (Ov). Assignment coverage is the fraction of documents assigned to clusters in relation with the total number of inputs (Co=(Res – Un) / Res) and overlap describes the fraction of documents confined to more than one group (Ov=(Nb / Res) –1). Fig. 7 shows these different values for each classifier.

Request	**Res.**	**AHC classification**					**STC classification**				
		Cl.	Nb	Un	Co	Ov	Cl.	Nb	Un	Co	Ov
Rice	55	37	50	5	0.9	-0.09	20	323	0	1	4.87
Animals	119	69	110	9	0.81	-0.08	20	755	5	0.958	5.34
Transport	28	18	28	0	1	0	14	84	14	0.5	2

Request	**Res.**	**LINGO classification**					**Metadata classification**				
		Cl.	Nb	Un	Co	Ov	Cl.	Nb	Un	Co	Ov
Rice	55	6	62	16	0.71	0.12	38	187	1	0.98	2.4
Animals	119	14	131	34	0.71	0.10	79	491	23	0.81	3.13
Transport	28	7	34	10	0.64	0.21	50	138	5	0.82	3.93

Request	**Res.**	**Combination**				
		Cl.	Nb	Un	Co	Ov
Rice	55	31	49	6	0.89	-0.11
Animals	119	82	116	3	0.97	-0.03
Transport	28	16	27	1	0.96	-0.04

Fig. 7. Analysis of different queries

We can observe that:

1. In AHC, the number of cluster is huge which implies a problem of legibility and quality of the heading of classes.
2. In STC, the documents are not sufficiently discriminating to deduce interesting regroupings, the overlap (Ov) is important.
3. In Lingo, there are less clusters, but there are more significant.
4. With the metadata, many clusters are detected because the thesaurus is very large.
5. With combination, the covertures of result is improved (almost complete).

Moreover, we observe that combination of these classifiers for the request “animals transport” highlight two important topics: “carcass transport between Europe and ex USSR” and “transport toward Spain” and minimize the title of the set of documents “publication”.

The combination gives prominence to the main topics and it reduces the individual errors of each classifier.

To simulate the user's contribution, we have defined 8 users shared in three groups: a customs officer group with two profiles "regulation" and "frontier", a veterinarian group ("sanitary", "disease", "butchery"), a farmer group ("cattleman", "fish breeding", "poultry farming"). For each user we add some keywords according to its concerns, for ex: "foot-and-mouth disease" for the veterinary in the 2nd group in charge of "disease".

We noticed that the cluster deduced from the user contribution is in strongly connection with his concerns (see Fig. 8).

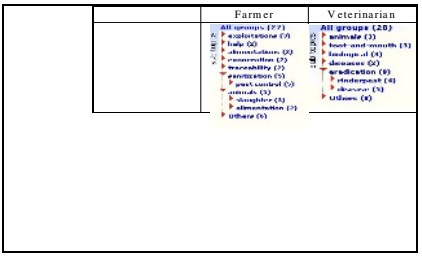

Fig. 8. User's categorization

At least, we obtain a better coverage and dispersion (number of clusters/ number of results) with the contributions of the group and the whole of the users (due to their complementary vision and a best knowledge on the documents). If we analyze the combination of all classifications, we continue to improve the coverage and the dispersion (see Fig. 9). For each user, the first group is the result of the individual annotations and the second is the result of the combination of the classifications.

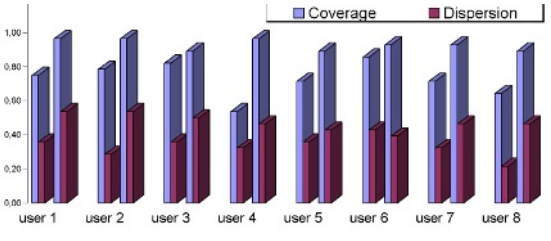

Fig. 9. Measure of coverage and dispersion for individual user's classification and combination of individual, group and all users classification

Fig. 10 shows the precision and recall for some classifications done on the request "animals transport". We can deduce that recall is strong on user's classifications. This implies a good segmentation. Recall and precision are more balanced with the implications of all users due to more view point on the document. A significant lower recall for the combination is observed. The information source is multiple thus implies less independence between classes. We obtain more classes so less independents.

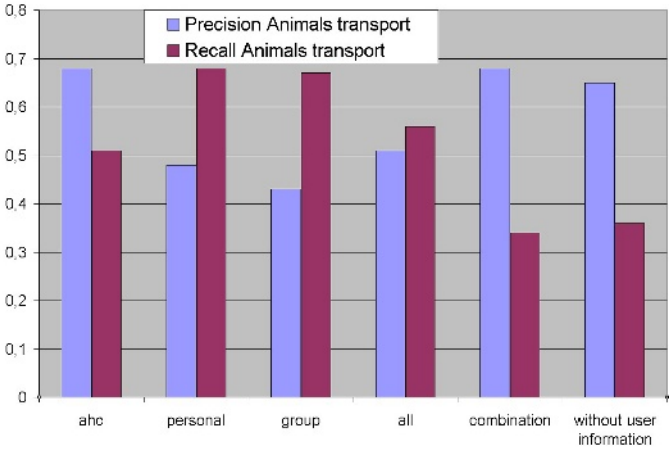

Fig. 10. Precision and recall

Of course, these analysis could be done to observe the influence of different parameters of classifiers. But that shows that they do not have a great impact. In fact, the test is only done with few documents, a hundred only.

However, that makes it possible to analyze the behaviour of each classifier and thus to measure that which is more in connection with the user's concerns.

All these tools and algorithms prepare the constitution of file, the reorganisation of the documents and the historic of them. Studies and measurements done to validate the various techniques allow to identify precise tasks and so to adapt to user's needs. The optimization of parameters and the understanding of each classifier give the highest performance on the system.

5 Conclusion

We have introduced a new system to help the user to build his own file according to his main interests. In order to facilitate the work of the user this system combines several ideas from the information retrieval domain: the sharing of knowledge on documents, the classification technologies and different tools to measure the efficiency of the algorithms. The novelty of the system is the usage of these possibilities drive by the user or recommended by the system.

The first experiments show that the various classification techniques allow a first regrouping of the documents. It is possible to refine this regrouping with the evaluation of the performances and the adaptations of algorithms. It is very useful to understand the possible proximities and relations between the documents. The user finds out more easily the content of the text with the different keywords. This could be a first step to explore the document and to order it in different folders. The second step is the appropriation of the document by the user. He can describe and annotate it. The contributions of different classes of users give more information on the document. Finally, the choice or the combination of these different approaches facilitate the creation of files.

In order to improve the system we have to add new classification algorithms. Working on document segments is another way to ameliorate the system. This work constitutes a first approach of file's constitution.

References

1. Rangoni, Y., Belaïd, A.: Data categorization for a context return applied to logical document structure recognition, ICDAR,Seoul, Korea (2005) 297-301
2. Hearst, M.A., Pedersen J.O.: Reexamining the cluster hypothesis: Scatter/Gather on retrieval results. In Actes of ACM/SIGIR Conference on Research and Development in Information Retrieval, Zurich, Suisse (1996) 76-84
3. Lam, W., Lai, K.Y.: A meta-learning approach for text categorization. In Proceedings of SIGIR-01,, New Orleans, US (2001) 303-309
4. Bennett, P.N., Dumais, S.T., Horvitz, E.: Probabilistic combination of text classifiers using reliability indicators: Models and results. In Proceedings of SIGIR-02, Tampere, Finland (2002) 207-215
5. Voorhees, E.M.: Implementing agglomerative hierarchical clustering algorithms for use in document retrieval, Information Processing and Management, vol. 22, (1986) 465-476,
6. Zamir, O., Etzioni, O.: Web document clustering: a feasibility demonstration, Proceedings of the 19th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98) (1998) 46-54
7. Osinski S.: An Algorithm for clustering of Web Search results", Master thesis, Poznan Universitv of technology (2003)
8. Lamirel, J.C., Francois, C., Al Shehadi, S., Hoffman M.: Multi-Topographic new classification quality estimators for analysis of documentary information: Application to patent analysis and web mapping. In Scientometrics international Journal, Vol. 60, No. 3 (2004) 445-462.