Cardinal , nominal or ordinal similarity measures in comparative evaluation of information retrieval process

Christine MICHEL

Laboratoire RECODOC
Université Claude Bernard Lyon 1
43 Bd du 11 novembre 1918, bat 721
69622 VILLEURBANNE Cedex
Tel : + 33 (0)4 72 43 13 91
Fax : + 33 (0)4 72 43 15 59

Laboratoire CEM-GRESIC
MSHA - Esplanade des Antilles, D.U
33607 PESSAC Cedex - FRANCE
Tél : +33 (0)5 56 84 68 13/ 68 14
Fax : +33 (0)5 56 84 68 10

Christine.Michel@montaigne.u-bordeaux.fr

Abstract

Similarity measures are used to quantify the resemblance of two sets. Simplest ones are calculated by ratios of the document's number of the compared sets. These measures are simple and usually employed in first steps of evaluation studies, they are called cardinal measures. Others measures compare sets upon the number of common documents they have. They are usually employed in quantitative information retrieval evaluations, some examples are Jaccard, Cosine, Recall or Precision. These measures are called nominal ones. There are more or less adapted in function of the richness of the information system's answer. Indeed, in the past, they were sufficient because answers given by systems were only composed by an unordered set of documents. But usual systems improve the quality or the visibility of there answers by using a relevant ranking or a clustering presentation of documents. In this case, similarity measures aren't adapted. In this paper we present some solutions in the case of totally ordered and partially ordered answer.

Introduction

The quantitative evaluation of a system's information retrieval process is often based upon the comparison of answers. For example, in large scale evaluation, system's answers are compared each to other (comparative evaluation) or to a referential set of "good" answer (diagnostic evaluation) (Hirschmann, 95). The calculus of the similarity between two answers C and C' depends on the richness of the information presentation format.

Indeed, basic information retrieval systems produce lists of documents without any particular order. Answer sets are in those cases compared in function of the number of documents they have (cardinal comparison) or the number of common documents they have (nominal comparison).

Actual information retrieval systems propose ranked list of documents as answer, the rank is given by the relevance degree from the document to the answer. It may be total or partial. For example, web engines give a completely ordered list of documents as answer, this is a **total order**. But many systems use the clustering process in order to improve the visibility of information set. *"Document clustering algorithms attempt to group documents together based on their similarities …/… This can help users both in location interesting document more easily and in getting an overview of the retrieved document set". "Information Retrieval community has long explored a number of post-retrieval document visualization techniques as alternatives to the ranked list presentation …/… document networks, spring embeddings, documents clustering, and self organizing map. Of the four major techniques, only document clustering appears to be both fast enough and intuitive enough to require little training or adjustment time from the user."* (Zamir, 99) In these case answer sets are **partially ordered**. Indeed, only clusters (i.e. classes) are ranked in order of relevance, documents are equally ranked in a cluster[1].

Searchers (Tague 1996) (Borlung 1998) quotes many studies which highlight the delay induced, in the satisfaction of an informational need, by a possible modification of this order of presentation. Ordinal measures must be used in order to take it into account in the calculus of the similarity, indeed, cardinal or nominal ones do not do it. But there are none really ordinal measures proposed for evaluation context. Indeed, measures proposed are most of time cardinal or nominal ones like Recall, Precision or Jaccard (Losee 90). The aim of this paper is to propose other ones.

In the first part of this paper we we'll describe the usual measures proposed in the case of evaluation tests: measures based upon nominal or cardinal comparisons. Then, in the second part, we'll present the total ordered formalism, the property the similarity measures must have in this case and examples of possible similarity measures. In the third and last part we'll present the most general case, the partial order. We'll explain why measures proposed in total order case can't be used there and how we can define new ordered similarity measures.

In the following section we will called :

D : a documents given as answer.

C and **C'** two sets of documents defining two answers to be compared.

[1] We can noticed that the partial order case is a generalization of the total order case; a total order is a partial order with clusters of one document.

1. No order case

1.1. Cardinal comparisons

Let suppose that there is not information upon documents of C and C'. The simplest possible comparison between C and C' may be made upon the number of elements they have. It is called the cardinal of sets C and C' and it's noted $|C|$ and $|C'|$. Corresponding indicators are ratios like:

$$\frac{|C|}{|C'|}, \frac{|C|}{|C|+|C'|} \text{ or } \frac{|C'|}{|C|+|C'|}$$

1.2. Nominal comparisons

Let's consider now that each document of C and C' is identified as a singular way i.e. like with a name. It's so possible to make nominal comparison, i.e. to identify the common documents of C and C'. The similarity between C and C' will grow with the number of common documents. They are mathematically represented by $C \cap C$ and graphically represented as in Figure 1.

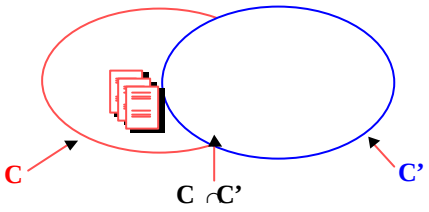

Figure 1

The number of common elements is $|C \cap C'|$.

Another important element is $|C \cup C'|$ which is the number of *different* documents of C and C' and calculated as : $|C \cup C'| = |C| + |C'| - |C \cap C'|$.

Most of usual similarity indicators are based upon this numbers. They differs from ones to other with the denominator number, calculated in order to normalize the measures from 0 to 1. As examples we can quotes (Boyce 94)(Losee 90):

The Jaccard's coefficient $\frac{|C \cap C'|}{|C \cup C'|}$ [Eq 1]

The Dice's coefficient $\frac{2\times|C \cap C'|}{|C|+|C'|}$ [Eq 2]

The cosine (Salton 83) $\frac{|C \cap C'|}{\sqrt{|C| \times |C'|}}$ [Eq 3]

The Overlap coefficient $\frac{|C \cap C'|}{\min(|C|,|C'|)}$ [Eq 4]

Measures of Recall and Precision used in the case of large scale evaluations or comparative test protocols like TREC (Voorhees 98) are also nominal ones.

2. The total order case

Let suppose now that the documents of C and C' are personalized by a name and presented in a totally ranked way. Let us call D_i and D'_j the documents of rank i and j from C and C'.

If C (respectively C') is composed of m (respectively m') documents we will have :

$C=\{D_1, D_2, \ldots, D_i, \ldots, D_m\}$ and
$C'=\{D'_1, D'_2, \ldots, D'_j, \ldots, D'_{m'}\}$.
The graphical representation is like Figure 2.

As before, the similarity indicators between C and C' will grow with the number of common documents. The rank give more detail about C and C' and permit to define other criteria. What are they?

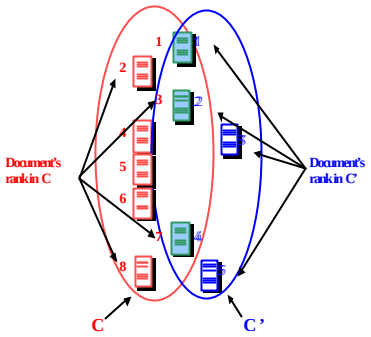

Figure 2

2.1. Possible similarity criteria in a ranking order case

2.1.1. The relative difference in order

Let suppose that D_i found as the rank i in C is the same document as D'_j found at the rank j in C'. The closest i and j are, the nearest C and C' should be and the higher the similarity should be. We'll call this criterion the relative difference in order. It represents the *difference* in the order of presentation of D_i *relative* to D'_j.

In the following example (Figure 3) (A, B) and (C, D) have identically one common document, nevertheless the similarity between A and B is higher than between C and D because the relative order difference of D_1 of A and D'_1 of B is less than the ones of D_2 of C and D'_1 of D.

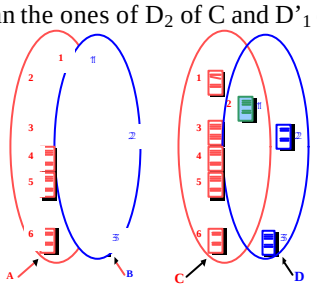

Figure 3

2.1.2. The top-ranking

Order of document are generally made upon a relevance criterion, so common documents presented at the end of the answer are less relevant than the ones presented in the beginning. This criterion must appear in similarity calculus, we'll call it *the top-ranking*. The more common documents are presented early to the users, the more similar the compared sets must be considered. As previous, let suppose that D_i and D'_j is the same document found at the rank i in C and at the rank j in C'. The higher i and j are, the smaller the similarity must be.

In the following example (Figure 4) the common documents D_2 of A and D'_1 of B are presented both previous to the user than D_5 of C and D'_2 of D. (A,B) and

(C,D) have both 1 document in common but, regards the top-ranking criterion, the similarity between C and D is higher than between E and F.

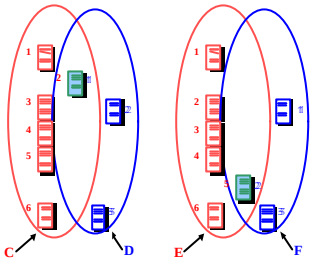

Figure 4

How construct similarity measures quantifying these criteria? The solution proposed by Tague (Tague 95) is to combine in the same similarity measure an indicator quantifying the similarity in terms (i.e varying with the number of common elements) with an indicator quantifying the similarity in order. This last one is called the delay indicator $\boldsymbol{d_0}$ and is increasing with the ordering difference of the set C to C'.

2.2. Delay : indicator of the order's difference between sets.

There exists many methods to construct delay indicators. Tague's idea is to calculated an adaptation of the rank correlation coefficient employed in statistic.

2.2.1. Delay derived from the coefficient correlation of rank

The correlation coefficient of rank quantifies the difference of two sets in terms of order. It's decreasing with the number of permutation used to rank the elements of C according to the elements of C'.

The adaptation presented by Tague (Tague 95) is called R and it's calculated like:

$$R(C,C') = \frac{\sum_{i=1}^{n} ir(D_i) - m(m+1)^2/4}{\sqrt{m(m^2-1)/12\left[\sum_{i=1}^{n} r(D_i)^2 - m(m+1)^2/4\right]}}$$

[Eq 5]

Where m is the total element number of C,

$r(D_i)$ the rank of the document D_i in C' if it's present.

She calculates the delay $\boldsymbol{d_0(R)}$ as a function of R.

This delay indicator is good because global but is not able to make appear the two previous criterions: the relative order and the freshness of information. Delay $\boldsymbol{a_{in}}$ and $\boldsymbol{m_{in}}$ presented below do it.

2.2.2. Delay calculated varying the relative order

The criterion defining the relative difference is : the closer i and j are, the higher the similarity must be and so the smaller the weight delay must be. So the similarity is decreasing with (i-j) or (j-i). The corresponding weight induce by the difference in order may be quantify by indicators like:

$$\boldsymbol{a}_{ij} = 1 - \frac{|i-j|}{\max(m,m')} \quad \text{[Eq 6]}, \qquad \boldsymbol{a}_{ij} = 1 - \frac{|i-j|}{nm'} \quad \text{[Eq 7]}$$

$$\boldsymbol{a}_{ij} = 1 - \frac{|i-j|}{n+m'} \quad \text{[Eq 8]}, \qquad \boldsymbol{a}_{ij} = \frac{1}{|i-j|} \quad \text{[Eq 9]}$$

2.2.3. Delay calculated varying the freshness of information

By using the same reasoning the similarity in terms of top-ranking is decreasing with the rank i and j. The delay induce may be quantify by indicators like:

$$\boldsymbol{m}_{ij} = 1 - \frac{|i\|j|}{(\max(n,m))^2} \quad \text{[Eq 10]}, \qquad \boldsymbol{m}_{ij} = \frac{(\max(n,m))^2}{|i\|j|} \quad \text{[Eq 11]}$$

$$\boldsymbol{m}_{ij} = 1 - \frac{|i\|j|}{nm} \quad \text{[Eq 12]}, \qquad \boldsymbol{m}_{ij} = \frac{nm}{|i\|j|} \quad \text{[Eq 13]}$$

$$\boldsymbol{m}_{ij} = 1 - \frac{|i\|j|}{n+m} \quad \text{[Eq 14]}, \qquad \boldsymbol{m}_{ij} = \frac{n+m}{|i\|j|} \quad \text{[Eq 15]}$$

2.3. Construction of ordered similarity measures

2.3.1. Type 1: Measures with general delay indication

Let's consider $S(C,C')$ as any nominal measures like Jaccard, Cosine, Dice, Recall, Precision previously describe in [Eq 1, 2, 3, 4]….

Tague construct it's similarity measure by using the indicator $\boldsymbol{d}_0(R)$ as in the further equation :

$$S_{1O}(C,C') = \boldsymbol{d}_0(R) S(C,C') \quad \text{[Eq 16]}$$

This calculus is possible because the indicator $\boldsymbol{d}_0(R)$ is global to C and C' and so is coherent with the other general similarity indication S(C,C'). Other global delay function may be calculated, for example with the means:

$$\overline{\boldsymbol{a}_{ij}} = \sum_{i=1}^{m}\sum_{j=1}^{m'} \frac{\boldsymbol{a}_{ij}}{m.m'} \quad \text{[Eq 17] or} \quad \overline{\boldsymbol{m}_{ij}} = \sum_{i=1}^{m}\sum_{j=1}^{m'} \frac{\boldsymbol{m}_{ij}}{m.m'} \quad \text{[Eq 18]}$$

In this case ordered similarity quantifying the relative difference in order may be:

$$S_{2O}(C,C') = \overline{\boldsymbol{a}_{ij}} S(C,C') \quad \text{[Eq 19]}$$

And the similarity quantifying the top-ranking difference may be :

$$S_{3O}(C,C') = \overline{\boldsymbol{m}_{ij}} S(C,C') \quad \text{[Eq 20]}$$

And it's also possible to combine delay like in the following formula:

$$S_{4O}(C,C') = \overline{\boldsymbol{a}_{ij}\,\boldsymbol{m}_{ij}} S(C,C') \quad \text{[Eq 21]}$$

2.3.2. Type 2: Measures with precise delay indication

Let suppose now that we want to link directly the criteria of relative difference or top-ranking to the concerned common documents. In this case, corresponding similarity measures can look like the following three measures :

$$S_{5O}(C,C') = \sum_{i=1}^{n}\sum_{j=1}^{m} \boldsymbol{a}_{ij} \left| D_i \cap D'_j \right| \quad \text{[Eq 22]}$$

$$S_{6O}(C,C') = \sum_{i=1}^{n}\sum_{j=1}^{m} \boldsymbol{m}_{ij} \left| D_i \cap D'_j \right| \quad \text{[Eq 23]}$$

$$S_{7O}(C,C') = \sum_{i=1}^{n}\sum_{j=1}^{m} \alpha_{ij}\, m_{ij} \left| D_i \cap D'_j \right| \qquad \text{[Eq 24]}$$

Type 1 measures are based upon a global indicator, which is not very precise. There is an opposite problem with type 2 measures : freshness of information or relative order is taken into account in a precise way but the comparison of the two sets elements is just made upon the intersection $\left|D_i \cap D'_j\right|$.

There is no solution to this dilemma in the total order formulation. Nevertheless, there is one if we consider the more general model of partial ordered sets presented below.

3. The partial order case

As we seen in introduction, systems tend to use clustering algorithms to improve the visibility of information. They produce classes of documents, classes are presented to the user in a ranking way, documents are usually equally ranked in a class. This is a partial order of documents graphically represented in Figure 5.

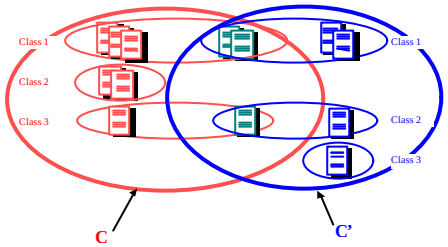

Figure 5

Let C_i and C'_i be the classes of rank i of the sets C and C', so $C=\{C_1,C_2, \ldots, C_i, \ldots, C_m\}$ and $C'=\{C'_1,C'_2, \ldots, C'_i, \ldots, C'_{m'}\}$.

Let's D_{ij} and D'_{ij} are the documents of sets C_i and C'_i. So $C_i=\{D_{11}, D_{12}, \ldots, D_{ij}, \ldots\}$, and $C'_i=\{D'_{11}, D'_{12}, \ldots, D'_{ij}, \ldots\}$.

In this case, the previous similarity measures (Eq 16, 19, 20, 21, 22, 23, 24) can't be adapted because of the classes formalism. Indeed :

- Let's remember that S(C,C') is a nominal measure like for example [Eq 1, 2, 3, 4]. If we considers the formula, S(C,C') can't be calculated without breaking the classes' hierarchy. In this case we totally loose the clustering information.
- The indices i and j haven't the same sense as before. It's convenient now to speak about $r(D_{ij})$, the rank of the document D_{ij} of C_i. The partial order hypothesis is that all the documents of a class have the same rank but what rank? Taking i as $r(D_{ij})$ is not the only possible solution.

The similarity measure $S(C,C')$ and the delay indicator must be adapted.

3.1.1. Possible adaptation of nominal similarity indicator

As we said before, the calculus of S(C,C') has no sense if we keep the classes formalism.

We advise to considered [Eq 25] as an indication of the sets C and C' nominal similarity.

$$S(C,C') = \sum_{i=1}^{m}\sum_{j=1}^{m'} S(C_i, C'_j) \qquad \text{[Eq 25]}$$

Indeed, in this case, C_i and C'_j are nominal sets and the classical similarity may be calculated.

3.1.2. Possible delays adaptation

In a partial order case, there is a dilemma in the choice of the rank $r(D_{ij})$ of the document D_{ij}. The simplest choice is : $r(D_{ij}) = i$ [Eq 26]

The formulas like [Eq 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15] aren't changed except in notation. For example [Eq 6] is written :

$$\alpha_{ij} = 1 - \frac{\left| r(D_{ij}) - r(D'_{kl}) \right|}{\max(m, m')} = 1 - \frac{|i-k|}{\max(m, m')} \qquad \text{[Eq 27]}$$

Let's considered now the case where the rank of the documents in a class depends on the documents class number. Let's call **m(j)** the element number of class C_j.

We can considered that the document D_j rank are determined with :

- the first element of the class :

$$r(D_{ij}) = \sum_{j=1}^{i-1} m(j) + 1 \qquad \text{[Eq 28]}$$

- the last element of the class :

$$r(D_{ij}) = \sum_{j=1}^{i} m(j) \qquad \text{[Eq 29]}$$

- The mean element of the class

$$r(D_{ij}) = \sum_{j=1}^{i-1} m(j) + \frac{m(i)}{2} \qquad \text{[Eq 30]}$$

The calculus of delay indicator are exactly made as in example in [Eq 27].Tague (Tague 95) advise to calculate $r(D_{ij})$ as the means rank ([Eq 30]).

It's possible now to adapt type 1 and type 2 similarity measure varying the chosen adaptations. But, the problems enunciated at the end of the section 2 are the same. Nevertheless, the formalism of partial ordered sets let's us think that there is another type of possible ordered similarity measures, measures of **type 3**.

3.1.3. Type 3 ordered similarity measures : combination of type 1 and 2.

Formalism of partial order case make the construction of ordered similarity measures as [Eq 31] possible.

$$Q(C,C') = \sum_{i=1}^{m}\sum_{j=1}^{m'} S(C_i, C'_j) \times \varphi(i,j) \qquad \text{[Eq 31]}$$

$\varphi(i,j)$ is defined by five particulars conditions[2] in order to make appears the relative order and the top-rank of information.

---

[2] (i) $\forall$ i, j$\in$ [1,m]x[1,m'], φ(i,j)>0

(ii) $\forall$ i$\in$ [1,m0], $\sum_{i=1}^{m_0} \varphi(i,i) = 1$

(iii) $\forall$ i, j$\in$ [1,m]x[1,m'], φ(i,j) is strictly decreasing in j$\geq$i (i fixed)

(iv) $\forall$ i, j$\in$ [1,m]x[1,m'], φ(i,j) is strictly decreasing in i$\geq$j (j fixed)

(v) $\forall$ i$\in$ [1,m$_0$] φ(i,i) is strictly decreasing in i.

We can noticed that information on similarity in document and in order delay are quantified in the more precise way as possible, i.e. for each class. So that's the reason why type 3 measure have both properties of type 1 and 2.

3.2. Example of type 3 measure

Let's suppose that $S(C_i,C'_j)$ is the Jaccard indicator ([Eq 1]), and $j\ (i,\ j)$ the combination of $d^{mo}(n)$ define as

$$j(i,\ j) = d^{mo}(i(|i-j|+1)) \times d^{mo}(j(|i-j|+1)) \qquad \text{[Eq 32]}$$

with : $$d^{mo}(n) = \frac{\sqrt{6m_0^{\ 3}}}{\sqrt{6m_0^{\ 4} - 6m_0^{\ 3} + 8m_0^{\ 2} - 3m_0 + 1}}\left(1 - \frac{n-1}{m_0^{\ 2}}\right)$$

and $m_0 = max(m,m')$

It's possible to construct similarity measure of type 3 like :

$$P_d(C,C') = \sum_{i=1}^{m} \sum_{j=1}^{m'} J(C_i, C'_j) \times d^{\ mo}(i(|i-j|+1)) \times d^{\ mo}(j(|i-j|+1))$$

[Eq 33]

The measure P_d has been tested in a real evaluation context : the diagnostic evaluation of a system having a personalized filtering process upon the user's profile. The aim and general methodology of the study is presented in (Michel 2000). A comparative study of the results given by a classical Jaccard measure and this one show that the delay induce by the relative order and the top ranking really justified the use of an ordered similarity measure (Michel 99).

Conclusion

Similarity measures are of three types : cardinal, nominal and ordinal ones. Cardinal ones are the simplest, they may be used in all the sets description case. Nominal ones are more precise if the sets have individual descriptions of all elements and ordinals one if, therefore, there is also an ordered classification of the elements. Sets of documents proposed as answer of information retrieval problems may be totally or partially ordered. In section 2 and 3 we propose some solutions and construct ordered similarity measure by combining two criterion : the similarity in terms of documents (called the *nominal similarity*) and the similarity in terms of order (called the *delay*). We can noticed that this two criterion can't be applied simultaneously in the total ordered case. Nevertheless, they can in the partial order case. This results from the measures construction choice, but also from the fact that, on reverse than in mathematics, the total order is, in the context of information retrieval, a particular case of the partial order.

References

(Borlung 98) : BORLUNG P., INGWERSEN P. – Measures of relative relevance and Ranked Half-Life : Performance indicators for interactive IR. – In *Proceeding of the SIGIR 98, 24-28 august 1998*, Melbourne, Australia

(Boyce 94) : BOYCE B.R., MEADOW C.T., KRAFT D.H. - *Measurement in information science*. – Academic Press – 1994 – 283 p.

(Ellis 96) : ELLIS D. – The dilemma of measurement in information retrieval research. - *Journal of the American Society for Information Science* – 47(1) - 1996 - pp 23-36.

(Frické 98) : FRICKE M – Jean Tague-Sutcliffe on measuring information – In *Journal of the American Society for Information Science* – 34(4) – 1998 – pp 385-394.

(Harter 96) : HARTER S.P. – Variation in Relevance assessment and measurement of retrieval effectiveness – In *Journal of the American Society for Information Science* – 47(1) - 1996 -pp 37-49.

(Harter 97) : HARTER S.P., HERT C.A. - Evaluation of information retrieval systems : Approches, Issues, and methods. In *Annual review of information science and technology* - 32 - 1997 -pp 1-93.

(Hirschman 95) : HIRSCHMAN L, THOMSON H.S., - Overview of evaluation in speech and Natural Langage Processing. In *Survey of the state of the art in human langage technology* collective direction COLE R./MARIANI J./USZKOREIT H./ZAENEN A./ZUE V. - Rapport NFS/CEE. To be published Cambridge University Press et Giardini Publ.

(Losee 90) : LOSEE R. M. – *The science of information. Measurement and applications* – Academic Press, Inc. – 1990 – 293 p.

(Michel 99): MICHEL C - Evaluation de systèmes de recherche d'information, comportant une fonctionnalité de filtrage, par des mesures endogènes. Réalisation et evaluation d'un prototype de système de recherche d'information avec filtre selon les profils des utilisateurs.- PhD Thesis - University Lyon II - 6 January 1999 -322 p.

(Michel 2000): MICHEL C. Diagnostic evaluation of a personalized filtering information retrieval system. Methodology and experimental results. In *Proceedings of RIAO 2000 "Content based multimedia information access"*, Collège de France, Paris, 12-14 april 2000

(Salton 83) : SALTON G. McGILL M. J. – *Introduction to modern Information Retrieval*. New York : McGraw-Hill.

(Tague 95) : TAGUE-SUTCLIFFE J. - *Mesuring information. An information services perspectives*. - Academic Press. - 1995 - 206 p.

(Tague 96) : TAGUE-SUTCLIFFE J. – Some perspectives on the evaluation of information retrieval systems In *Journal of the American Society for information science*. - 47(1) - 1996 - pp 1-3.

(Voorhees 98) : VOORHEES E.M. – HARMAN D. – Overview of the seventh Text Retrieval Conference TREC 7. In *Proceedings of the seventh Text Retrieval Conference TREC 7*. – Gaitherburg 9-11 nivember 1998 - p.

(Zamir 99) : ZAMIR O., ETZIONI O. : Grouper : A dynamic clustering interface to web search results – In *Proceeding of the Eighth International World Wide Web Conference* – May 11-14 1999 – Toronto, Canada (http://www8.org)