Order-based Equivalence Degrees for Similarity and Distance Measures

Marie-Jeanne Lesot, Maria Rifqi

Order-based Equivalence Degrees for Similarity and Distance Measures

Marie-Jeanne Lesot and Maria Rifqi

Université Pierre et Marie Curie - Paris 6, CNRS, UMR7606, LIP6
104 avenue du President Kennedy, 75016 Paris, France
{marie-jeanne.lesot,maria.rifqi}@lip6.fr

Abstract. In order to help to choose similarity or distance measures for information retrieval systems, we compare the orders these measures induce and quantify their agreement by a *degree of equivalence*. We both consider measures dedicated to binary and numerical data, carrying out experiments both on artificial and real data sets, and identifying equivalent as well as quasi-equivalent measures that can be considered as redundant in the information retrieval framework.

Key words: similarity, distance, kernel, order-based comparison, equivalence degree, Kendall tau

1 Introduction

Information retrieval systems provide results in the form of document lists ordered by relevance, usually computed as the similarity between the document and the user request. The choice of the similarity measure is then a central component of the system. In such applications, the similarity values themselves are of little importance, only the order they induce matters: two measures leading to the same document ordering can be considered as equivalent, and it is not useful to keep them both. Likewise, several machine learning algorithms only depend on the similarity rankings and not on their values, such as the k-nearest neighbor classification, hierarchical clustering with complete or single linkage, or the monotone equivariant cluster analysis [1].

To formalize this notion, several authors introduced the definition of *equivalent* comparison measures [2–5], as measures always inducing the same ranking, and exhibited classes of equivalent measures. To refine the characterization of non-equivalent measures, *equivalence degrees* were then proposed [6] to quantify the disagreement between the rankings, considering both the number of inversions and their positions, through the generalised Kendall tau [7, 8].

In this paper we propose a systematic study of these equivalence and quasi-equivalence properties both for measures dedicated to presence/absence and to numerical data, i.e. data respectively in $\{0,1\}^p$ and in $\mathbb{R}^p$, taking into account the main existing similarity, distance and scalar product measures. We compute the equivalence degrees considering both artificial and real data, the latter consisting of training data from the 2008 Image CLEF challenge [9].

As opposed to previous work [6], the protocol we consider here corresponds to the use of an information retrieval system: it consists in comparing to a request data all n points of the data set, ranking them according to their similarity to this request and averaging the result over several requests. This better reflects the application case, whereas the protocol used in [6] considering all $n(n-1)/2$ data pairs simultaneously and ordering them in a single ranking was more focused on a theoretical comparison of similarity measures. Furthermore, in this paper, we extend the comparison framework to the case of numerical data.

The paper is organised as follows: section 2 recalls the definitions of equivalence and equivalence degrees for comparison measures and details the experimental protocol. Sections 3 and 4 respectively analyse the results obtained in the case of binary and numerical data.

2 Order-based Comparison of Comparison Measures

Denoting $\mathcal{X}$ the data universe, similarity measures are functions $S : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ quantifying proximity or resemblance: they take as argument object couples and give as a result numerical values that are all the higher as the objects are close. Distance measures $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ quantify dissimilarity and return values that are all the smaller as the objects are close. Similarity and distance measures build the set of comparison measures.

2.1 Definitions

Order-based Equivalence Several authors [2–5] considered the issue of a theoretical comparison between similarity measures and defined two measures m_1 and m_2 as *equivalent* if they induce the same order when comparing objects: more formally they are equivalent if and only if $\forall x, y, z, t$, it holds that $m_1(x, y) < m_1(z, t) \Leftrightarrow m_2(x, y) < m_2(z, t)$ and $m_1(x, y) = m_1(z, t) \Leftrightarrow m_2(x, y) = m_2(z, t)$.

It has been shown [4, 5] that, equivalently, m_1 and m_2 are equivalent if and only if there exists a strictly increasing function $f : Im(m_1) \to Im(m_2)$ such that $m_2 = f \circ m_1$, where $Im(m) = \{s \in [0, 1]/\exists (x, y) \in \mathcal{X}^2, s = m(x, y)\}$.

It is to be noted that when a distance is compared to a similarity measure, it is necessary to take into account their opposite sense of variation: the inequalities in the first definition must be the opposite one of the other; the function of the second definition must be strictly decreasing.

Order-based Equivalence Degrees In order to refine the characterization of non-equivalent measures, it has been proposed to quantify the disagreement between the induced rankings, by *equivalence degrees* [6]: two measures leading to a few inversions can be considered as more equivalent than measures inducing opposite rankings. Furthermore, two measures can be considered as less equivalent if the inversions occur for high similarity values than if they occur for low values: in the framework of information retrieval systems for instance, most often

only the first results are taken into account, inversions occurring at the end of the document lists are not even noticed.

The generalized Kendall tau K_{p_t,p_m} [7, 8] compares two rankings r_1 and r_2 defined on a set of elements $\mathcal{E}$, taking into account the number of inversions as well as their positions: it associates each element pair $(i, j) \in \mathcal{E}^2$ with a penalty $P(i, j)$ and is defined as the sum of all penalties divided by the number of pairs. Four penalty values are distinguished: if the pair (i, j) is concordant (i.e. r_1 and r_2 agree on the relative position of i and j: formally denoting $\delta_l = r_l(i) - r_l(j)$ the rank difference of i and j in ranking r_l, if $\delta_1\delta_2 > 0$ or $\delta_1 = \delta_2 = 0$), then $P = 0$; if the pair is discordant (i.e. $\delta_1\delta_2 < 0$), $P = 1$; if it is tied in one ranking but not in the other one, $P = p_t \in [0, 1]$. Lastly if it is present in one ranking but missing from the other one, one distinguishes whether both i and j are missing ($P = p_m \in [0, 1]$), or only one is (the pair is then handled as a normal one).

The equivalence degree between two comparison measures m_1 and m_2 is thus computed as follows: given a data set $\mathcal{D}$ and a request $x \in \mathcal{D}$, all points $y \in \mathcal{D}$ are ranked according to their similarity to x, according to m_1 and m_2. The rankings r_1^k and r_2^k induced on $\mathcal{D}$, restricted to their top-k elements, i.e. to the objects with rank smaller than a given k are then compared, leading to:

$$d_{\mathcal{D}}^k(m_1, m_2) = 1 - K_{0.5,1}(r_1^k, r_2^k)$$

It equals 1 for equivalent measures and 0 for measures leading to opposite rankings. We set $p_t = 0.5$ considering that when breaking a tie, there is 1 chance out of 2 to come up with the same order as defined by the second ranking. We set $p_m = 1$ considering that a missing data pair indicates a major difference and can be penalized as a discordant pair. Lastly, for any given k, each data point $x \in \mathcal{D}$ is successively considered as request, and the degrees are averaged over all requests.

2.2 Considered Data Sets

We carry out experiments considering both binary and numerical data, i.e. respectively the universes $\mathcal{X} = \{0, 1\}^p$ and $\mathcal{X} = \mathbb{R}^p$, and for each of these two types, artificial and real data set.

For the real data, we consider the ImageClef training corpus [9] that contains 1827 images annotated in a multi-label framework (e.g. indicating whether the image shows buildings or vegetation). On one hand we use the image labels to define binary data, encoding the presence or absence of each label. We suppressed some labels in XOR relation with others (such as night, related to day, or outdoor, related to indoor) as well as subcategory labels (tree, subsumed by vegetation, and sunny, partly cloudy and overcast subsumed by sky). As a result, the binary data set contains $p = 11$ attributes. On the other hand, we encode the images using their histograms in the HSV space (using $p = 6 \times 2 \times 2 = 24$ bins) expressed as percentages, to get a vector description. It is to be noted that this vector description is such that the sum of all attributes is constant.

The artificial data are generated according to the real data, so as to study the effect on equivalence results of potential specific data configurations, e.g. variable

Table 1. Classic binary data similarity measures, normalised to $[0, 1]$ (the definitions may thus differ from the classic ones).

Similarity measure	*Notation*	*Definition*
Jaccard	Jac	$\frac{a}{a+b+c}$
Dice	Dic	$\frac{2a}{2a+b+c}$
Kulczynski 2	Kul	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$
Ochiai	Och	$\frac{a}{\sqrt{a+b}\sqrt{a+c}}$
Rogers and Tanimoto	RT	$\frac{a+d}{a+2(b+c)+d}$
Russel and Rao	RR	$\frac{a}{a+b+c+d}$
Simple Matching	SM	$\frac{a+d}{a+b+c+d}$
Sokal and Sneath 1	$SS1$	$\frac{a+d}{a+\frac{1}{2}(b+c)+d}$
Yule Q	YuQ	$\frac{ad}{ad+bc}$
Yule Y	YuY	$\frac{\sqrt{ad}}{\sqrt{ad}+\sqrt{bc}}$

density or cluster structures. In the binary case, the artificial data consists of all points in a regular grid in $\{0, 1\}^{11}$, resulting in $2^{11} = 2048$ points. In the numerical case, the artificial data set is randomly generated following a uniform distribution on $[0, 100]^{24}$.

3 Binary Data Similarity Measures

3.1 List of Considered Measures

Formally, similarity measures for binary data are defined as functions $S : \{0, 1\}^p \times \{0, 1\}^p \to \mathbb{R}$ possessing the properties of maximality ($\forall a, y,\ S(x, x) \geq S(x, y)$) and symmetry [10, 11], although the latter is not always required [12].

Table 1 recalls the definition of 10 classic similarity measures, using the following notations: for any point $x \in \{0, 1\}^p$, X denotes the set of attributes present in x, i.e. $X = \{i | x_i = 1\}$; for any data pair (x, y), a, b, c, d denote the number of attributes respectively common to both points $a = |X \cap Y|$, present in x but not in y or vice-versa, $b = |X - Y|$ and $c = |Y - X|$, and present in neither x nor y, $d = |\bar{X} \cap \bar{Y}|$. The measures not depending on d (the first 4 in Table 1) are called *type I similarity measures*, the others *type II similarity measures*. As can be seen from the table, the first 2 measures follow the same general scheme proposed by Tversky [12] $\mathrm{Tve}_{\alpha,\beta}(x, y) = a/(a + \alpha b + \beta c)$ corresponding to the special case where $\alpha = \beta = 1$ or $1/2$ respectively.

3.2 Analytical Equivalence Results

Several classes of equivalent similarity measures were established, exhibiting their functional dependency [3–5]. For the measures defined in Table 1 they are: (i) {Jaccard, Dice, symmetrical Tversky's measures $\mathrm{Tve}_{\alpha,\alpha}$}, (ii) {Rogers and Tanimoto, Simple Matching, Sokal and Sneath 1}, (iii) {Yule Q, Yule Y},

Table 2. Full rank equivalence degrees for artificial binary data.

	Jac	Kul2	Och	RT	RR	SM	SS1	YuQ	YuY	Random
Dic	1	0.97	0.99	0.87	0.89	0.87	0.87	0.86	0.86	0.50
Jac		0.97	0.99	0.87	0.89	0.87	0.87	0.86	0.86	0.50
Kul2			0.98	0.88	0.88	0.88	0.88	0.88	0.88	0.50
Och				0.88	0.89	0.88	0.88	0.87	0.87	0.50
RT					0.76	1	1	0.90	0.90	0.50
RR						0.76	0.76	0.77	0.77	0.50
SM							1	0.90	0.90	0.50
SS1								0.90	0.90	0.50
YuQ									1	0.50
YuY										0.50

(iv) each of the remaining measures forming a class by itself. For the Tversky's measures, it was more generally shown [5] that two measures with parameters (α, β) and (α', β') are equivalent if and only if $\alpha/\beta = \alpha'/\beta'$.

3.3 Experimental Results

Full Rank Comparison Table 2 contains the full rank equivalence degrees computed in the case of the artificial data. The top graph of Figure 1 offers a graphical representation of these values, together with their standard deviation.

As a baseline, we include a measure that generates random similarity values so as to have a reference equivalence degree. This measure has an equivalence degree of 0.5 with all measures: on average it ranks differently half of the pairs.

From the equivalence degrees equal to 1, three groups of equivalent measures are numerically identified, accordingly to the theoretical results (see Section 3.2). The non-1 degrees give information on the non equivalent measures. It can first be noted that they all have high equivalence levels: apart from the random measure, the minimal degree equals 0.76, which implies that the proportion of inversions is always lower than 24%. Furthermore, it appears that some measures, although not satisfying the definition of equivalence, have very high equivalence degrees, above 0.97 (Jac/Och, Kul2/Och, and Jac/Kul2): the latter, that actually equals the set of type I measures, lead to very few differences and can actually be considered as quasi-equivalent and thus redundant.

Figure 1 illustrates these degrees with their standard deviation, representing measure pairs in decreasing order of their degrees. To improve the readability, it only represents a single member of each equivalence class, and does not consider further the random measure. Taking into account the standard deviation, it can be observed on the top graph that for full rank comparison there is no significant difference between the degrees computed on the artificial and the real data. Thus all comments on the measures also hold for the real data set.

This graph also highlights the difference between the two measure types, as already mentioned: whereas type I measures appear highly equivalent one to another, the "intra equivalence" of type II measures is smaller. The latter do

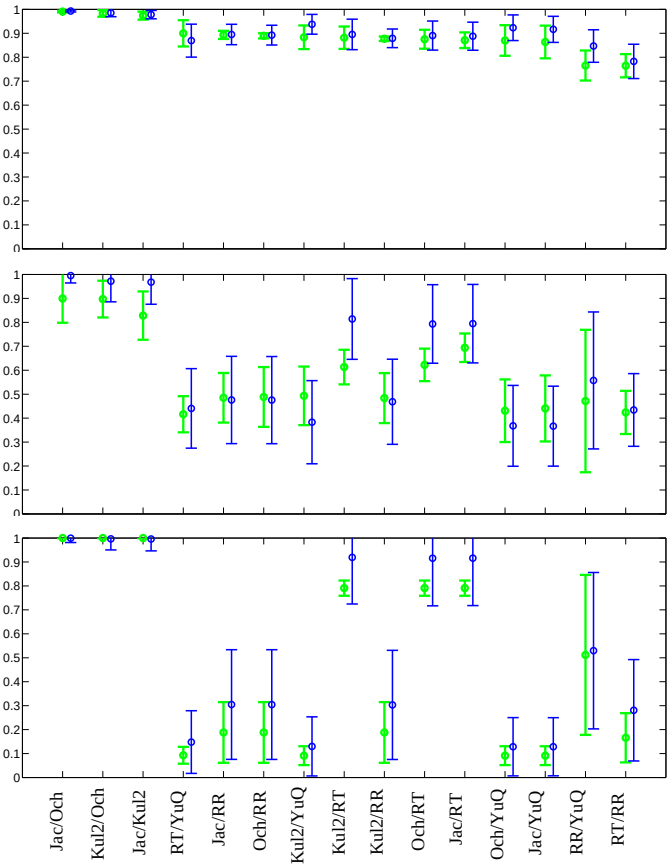

Fig. 1. Equivalence degrees and their standard deviation: (top) full ranking, (middle) top-100 (bottom) top-10. For each measure pair the left (resp. right) bar corresponds to artificial (resp. real) data.

not resemble each other more than they resemble the type I measures, which makes their category less homogeneous and more diverse.

Top-k Comparison The middle and bottom graphs of figure 1 show the equivalence degrees obtained when considering, respectively, the top-100 and top-10 ranked lists. We keep the same abscisse axis used for the full ranking, to underline the differences occuring when the list is shortened.

It can first be observed that the degrees are globally lower than for the full rank comparison: the minimum is 0.42 for $k = 100$, 0.09 for $k = 10$, indicating major differences in the ranked lists provided by the measures. The equivalence degree of the random measure with any other one (not shown on the graphs)

falls down below 0.1: the list it induces has next to nothing in common with the other lists, and almost all data pairs get a missing penalty.

This decrease indicates that the global agreement observed when comparing the full rankings is actually mainly due to the last ranked data. This underlines that a study of the inversion positions, besides their number, is necessary, especially when it comes to selecting non equivalent measures in an information retrieval framework. Still, this decrease does not occur for all measures: the intra type I pairs as well as those involving a type I measure with Rogers Tanimoto appear to be stable from full ranking to top-100 and top-10. Due to this behaviour, RT, although being a type II measure, is closer to the type I category than to type II. These measures can be considered as equivalent even for restricted rankings, and redundant for information retrieval applications.

Another difference when focusing on the top-k rankings comes from the standard deviations: it appears that their values considerably increase. Furthermore, they globally take higher values on the real data than on the artificial ones. This may be due to the regular distribution of the artificial data, which insures independence with respect to the request data. On the contrary, the real data probably follow a distribution with variable density, and the data request may have different effects, depending on whether it belongs to a dense or to a sparse region. Still, as for the full rank comparison, and except for RT, no significant difference between artificial and real data can be observed.

Lastly, it appears that the Yule Q and Russel Rao measures become the most isolated ones, far from all others: for YuQ, this can be explained by the fact that it very often takes value 1. Indeed, this occurs for all data pairs (x, y) such that $b = 0$ or $c = 0$. Thus, the set of data in its top-k list is much larger than those of the other measures, leading to many missing data pairs. The RR behaviour can be explained similarly: this measure only takes $p + 1 = 12$ different values in a universe of size p. Thus its top-k lists contain the whole data set even for low k values, again leading to many missing pairs when comparing to other measures.

4 Numerical Data Similarity Measures

4.1 List of Considered Measures

Numerical data comparison measures are based on distances or on scalar products [11]. The formers possess properties of positivity, symmetry, minimality, equivalently to the binary data similarity measures. Moreover, they satisfy the triangular inequality. The most classic distances are the Minkowski family, and in particular the Euclidean distance, denoted d_e, and the Manhattan distance.

The most common dot products comprise the Euclidean dot product k_e, the gaussian kernel $kg_\sigma = \exp(-d_e(x, y)^2/(2\sigma^2))$ and the polynomial kernel $kp_{\gamma,l} = (\langle x, y\rangle + l)^\gamma$. With the exception of the gaussian kernel, they do not correspond to classic similarity measures because they do not possess the maximality property, as e.g. $k(x, 2x) > k(x, x)$. To obtain it, it is necessary to normalize them, defining $\tilde{k}(x, y) = k(x, y)/\sqrt{k(x, x)k(y, y)}$. The similarity then only depends on the angle between the two vectors.

Table 3. Full rank equivalence degrees for artificial numerical data.

	L2	EDP	NEDP	GK50	GK100	PK3	NPK3	Random
L1	0.90	0.63	0.84	0.90	0.90	0.63	0.89	0.50
L2		0.63	0.87	1	1	0.63	0.97	0.50
EDP			0.76	0.63	0.63	1	0.66	0.50
NEDP				0.87	0.87	0.76	0.90	0.50
GK50					1	0.63	0.97	0.50
GK100						0.63	0.97	0.50
PK3							0.66	0.50
NPK3								0.50

4.2 Analytical Results

Using the functional definition of equivalence, two equivalence classes can be distinguished. The first one obviously groups the Gaussian kernels with the Euclidean distance: $kg_\sigma = f \circ d$ with $f : x \mapsto \exp(-x^2/(2\sigma^2))$ that is decreasing. All Gaussian kernels are thus equivalent: in particular, this implies that all σ values always lead to the same ranking.

The second class, grouping the Euclidean dot product and the polynomial kernels, is defined down to a data translation: for even values of γ, the function $g(x) = (x + l)^\gamma$, such that $kp_{\gamma,l} = g \circ k_e$, is increasing only under the condition that $x \geq -l$. Now denoting α the value such that $\forall x \forall i\, x_i + \alpha \geq 0$ and e the vector such that $\forall i\ e_i = \alpha$, after applying the translation by e, one has $\forall x \forall i\ x_i \geq 0$ and thus $\langle x, y \rangle = \sum_i x_i y_i \geq 0 > -l$. It can be underlined that in a classification framework the l value does not matter as it scales the feature space attributes and is counterbalanced by the weighting coefficient learned by the classifier.

In the case where the data are such that $\|x\| = 1$ for all x, these two classes are merged: indeed $d_e = h \circ k_e$ with $h(x) = \sqrt{2(1 - x)}$ that is strictly decreasing.

4.3 Experimental Results

We compare the most common measures namely the Manhattan (denoted L1) and Euclidean (L2) distances, the Euclidean dot product (EDP) and its normalised form (EDPN), the Gaussian kernel for $\sigma = 50$ (GK50) et $\sigma = 100$ (GK100), the polynomial kernel of degree 3 for $l = 2000$ (PK3) and its normalisation (NPK3). The σ and l values for the GK and PK were chosen according to the data properties. We also add a baseline random measure.

Full Rank Comparison Table 3 contains the full rank equivalence degrees, also illustrated, together with their standard deviation, on the top graph of figure 2.

As in the binary data case, and for the same reason, the random measure has an equivalence degree of 0.5 with all measures. The degrees equaling 1 are concordant with the theoretical results and indicate the two expected equivalence classes. Again, all measures have a high agreement level, as the maximal

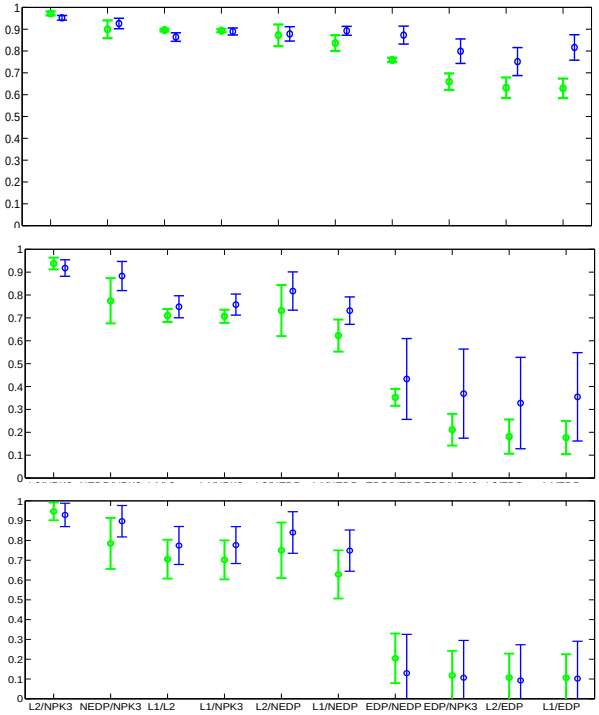

Fig. 2. Equivalence degrees and their standard deviation, for artificial and real numerical data.

proportion of inversions is only 37%, obtained when comparing the Gaussian and polynomial kernels. The observed high degree between L2 and NPK3 does not correspond to a theoretically known result. It can be explained by the level lines of these measures (figure omitted for space constraints): even if they locally differ, they have the same global form and the orders they induce globally agree.

The top graph of figure 2 highlights a difference between the artificial and real data sets that leads to a slightly different ordering of the measure pairs according to their equivalence degrees. This can be explained by the particularity of the real data: as they correspond to repartition histograms, their L1 norm is constant. This specific structure of the data has consequences on the equivalence degrees.

Top-k Comparison When focusing on top-k rankings, it can be observed that the difference between the two data types becomes less marked when k decreases. The standard deviations increase, underlying the influence of the request data especially on the beginning of the lists. Besides, although the equivalence degrees significantly decrease, the order of the measure pairs in terms of equivalence degree is not modified. Three equivalence levels can be distinguished in particular for $k = 10$. The highest one is reached by the pair L2/NPK3, meaning that their

high agreement holds for the highest similarity values. The lowest values are reached by EDP and any other measures: EDP appears as an isolated measure which has very less in common with the rest of the measures.

5 Conclusion

We compared similarity measures for two different data types, quantifying their proximity and possible redundancy when looking at the ranking they induce, and considering in particular restricted rankings associated to top-k lists. This study, relying on the definition of equivalence degree based on the generalised Kendall tau, takes place in the framework of information retrieval systems. Carrying out experiments both on artificial and real data, we showed some stability property regarding the behaviors of comparison measures on equivalence and quasi-equivalence results, but also some differences confirming that the equivalence degrees depends on the data sets but less than one could expect.

In future works, we aim to establish relations between data set structure and quasi-equivalence classes of measures of similarity. Lerman [2] considered this point of view in the case of binary data, showing that if all data have the same number of present attributes, i.e. if $\exists q / \forall x \in \mathcal{D}\, |X| = q$, then all similarity measures are equivalent on $\mathcal{D}$. We would like to extent this study to numerical data and to the quasi-equivalence property.

References

1. Janowitz, M.F.: Monotone equivariant cluster analysis. SIAM J. Appl. Math. **37** (1979) 148–165
2. Lerman, I.C.: Indice de similarité et préordonnance associée. In: Séminaire sur les ordres totaux finis, Aix-en-Provence (1967) 233–243
3. Baulieu, F.B.: A classification of presence/absence based dissimilarity coefficients. Journal of Classification **6** (1989) 233–246
4. Batagelj, V., Bren, M.: Comparing resemblance measures. Journal of Classification **12** (1995) 73–90
5. Omhover, J.F., Rifqi, M., Detyniecki, M.: Ranking invariance based on similarity measures in document retrieval. In: Adaptive Multimedia Retrieval AMR'05. Springer LNCS (2006) 55–64
6. Rifqi, M., Lesot, M.J., Detyniecki, M.: Fuzzy order-equivalence for similarity measures. In: Proc. of NAFIPS 2008. (2008)
7. Fagin, R., Kumar, R., Sivakumar, D.: Comparing top k lists. SIAM Journal on Discrete Mathematics **17** (2003) 134–160
8. Fagin, R., Kumar, R., Mahdian, M., Sivakumar, D., Vee, E.: Comparing and aggregating rankings with ties. In: Symp. on Princ. of Database Sys. (2004) 47–58
9. ImageCLEF challenge: http://www.imageclef.org (2008)
10. Bouchon-Meunier, B., Rifqi, M., Bothorel, S.: Towards general measures of comparison of objects. Fuzzy sets and systems **84** (1996) 143–153
11. Lesot, M.J., Rifqi, M., Benhadda, H.: Similarity measures for binary and numerical data: a survey. Intern. J. of Knowledge Engineering and Soft Data Paradigms (KESDP) **1** (2009) 63–84
12. Tversky, A.: Features of similarity. Psychological Review **84** (1977) 327–352