Novelty and Diversity in Image Retrieval

Simone Santini
Universidad Autónoma de Madrid, Spain

Keywords: Novelty, Diversity, Redundancy in Query Results, Evaluation.

Abstract: This paper studies the formalization and the use of the concepts of *novelty* and *diversity* to diversify the result set of a multimedia query, avoiding the presence of uninformative results. First, we review and adapt several diversity measures proposed in the information retrieval literature. The problem of maximizing diversity being NP-complete, we propose a general greedy algorithm (dependent on a scoring function) for finding an approximate solution, and instantiate it using three scenarios: a probabilistic one, a fuzzy one, and a geometric one. Finally, we perform tests on two data sets, one in which retrieval is based on annotations and the other in which retrieval is purely visual.

1 INTRODUCTION

Consider a multimedia data base D, with $|D|$ items to which a query q is submitted. A standard retrieval system will assign to each item $d \in D$ a *relevance* value for the query q, $r(d|q)$, and, assuming that the output of the system consists of a list with n slots, the system will show the results $[d_1, \ldots, d_n]$ with $r(d_1|q) \geq r(d_2|q) \geq \cdots \geq r(d_n|q)$ and $r(d_n|q) \geq r(d_k|q)$ for $k > n$.

The origin of this model, often called the *Robertsonian* model of relevance, is in information retrieval, in particular in (Robertson and Spark-Jones, 1976). Despite its rather neutral and straightforward appearance, the Robertsonian model is based on a number of fairly strong assumptions about the nature of relevance (Saracevic, 2007). One of these assumptions, in which we are specifically interested here, is that of *independence*: Robertson assumes that relevance is a *property* of an item vis-à-vis the query, and it does not depend on the relevance of other items in the result set. Around the turn of the XXI century, Information Retrieval researchers began to question this assumption (see, e.g. the aforementioned (Saracevic, 2007)). The accusation that was moved to it is that it may lead to result sets formally correct but not very informative. In some data bases there are a lot of very similar items that contain more or less the same information; if one of them is very relevant for a query, it is likely that all of them will be, and that the result set will be composed of items very much alike. This is true especially in the age of the internet, in which any conceivable information is repeated manifold. Although formally relevant, each one of these items adds very little information to what one already has with just one of them. In multimedia, this translates to situations like that of figure 1, which shows

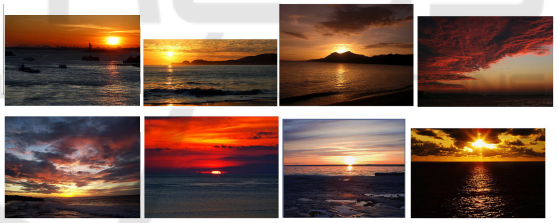

Figure 1: A result corresponding to query by example in which nearly all images contain the same information.

the results of a query by example, similarity based search engine. The first image (top, left) is the query. The results are formally correct under a Robertsonian interpretation, since all the images are very similar to the query and hence very relevant. Nevertheless, the images are so much alike as to be virtually interchangeable, and the whole set is quite poor from the point of view of the information provided to the user: there is little variety for the user to choose from, and no good idea of what alternatives the data base has to offer to satisfy the user's needs. In multimedia, the independence assumption presents an additional risk due to the inherent imprecision of the methods used to estimate relevance. Figure 2 shows the results of the query "car" executed on the annotation data base *Im2Text* (Ordoñez et al., 2011) using standard information retrieval techniques. While, at first sight, these

Santini, S.
Novelty and Diversity in Image Retrieval.
DOI: 10.5220/0006856703370346
In *Proceedings of the 15th International Joint Conference on e-Business and Telecommunications (ICETE 2018) - Volume 1: DCNET, ICE-B, OPTICS, SIGMAP and WINSYS*, pages 337-346
ISBN: 978-989-758-319-3

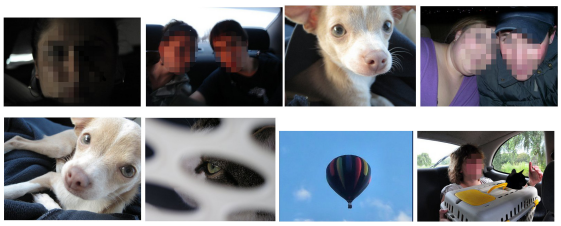

Figure 2: A result corresponding to the query *car* on an annotated data base.

images seem to have little relevance for the query, they all had captions such as "in the car," "in our car," "seen from my car," and so on. For all of them, after stop-word removal and stemming, the only stem left was *CAR* which therefore received, in the normalized vector model, a weight of 1. These images had therefore maximal similarity with the query and, in the absence of provisions for diversifying the results, the system guilelessly put them in the first place. The system here is stuck in a "semantic rut:" a lot of images share the same haphazard characteristic that makes them falsely suitable for the query and, given the frequency of occurrence of this trait, hoard the first positions of the result list.

These examples are not a slapdash collection of fortuitous cases, but a representative of a general phenomenon: relevance, alone, is not a suitable basis for satisfactory retrieval. In information retrieval, the Robertsonian hypothesis that relevance is a property of a document has all but been abandoned. It is, we argue, time that the multimedia community follow suit.

In the multimedia community, attempts have been made to solve the problems caused by examples such as that of figure 1 through *near-duplicate elimination*. Unfortunately, these inchoate techniques do not relate near-duplicity to the information content of the result set. Consequently, they fail to take a global view of the result set, to consider it as a mathematical object with a precise function, a function that can be invalidated by elements that are not necessarily duplicates. This makes near-duplicate elimination of limited usefulness in annotation-based systems (*vide* the example of figure 2, in which no near-duplicates are present), or on hybrid similarity-annotation systems (Rasiwasia et al., 2010).

In this paper, we propose a *non-Robertsonian* framework to deal with these issues, one based on the notions of *novelty* and *diversity*, with which Information Retrieval conceptualizes the problems caused by the independence hypothesis. These concepts, and the measures that come with them, will allow, on the one hand, to avoid the limitations of near-duplicate removal and, on the other hand, the formulation of a coherent theory that will apply to visual similarity systems, annotation-based systems, and their hybrids.

Diversity is the notion that allows the result set to deal with query *ambiguity*, while novelty deals with query *underspecification*. Consider a query composed of the keyword *manhattan*. The query is ambiguous as it can have several *interpretations*: it may refer to one of the "boroughs" of New York, to the cocktail, to the Woody Allen movie, or to the Indian tribe from which the Dutch bought the island. A result set with high diversity will cover all these interpretations, possibly in a measure proportional to an *a priori* estimation of the interest in each one of them. If we concentrate on a single interpretation (say: the borough) there are many different *aspects* in which one may be interested. We may be interested in the history of Manhattan, in its attractions, in getting around it, or in the housing prices. While interpretations are assumed to be mutually exclusive (if I am interested in the movie, I am probably not interested in the Indian tribe), aspects are inclusive: I am more or less interested in all of them. An item in a result set is *novel* to the extent in which it covers aspects of a query not covered by other items in the result set, that is, to the extent in which items are *non-redundant*: removing an item would lead to a result set that would not cover one or more of the aspects covered by the set before the removal. Diversity is a global property of a data set, while novelty is the corresponding property of a document with respect to a set.

In the last few years, various methods have been proposed both to measure the diversity and novelty of a set of items (Chapelle et al., 2009; Clarke et al., 2009; Santini and Castells, 2011) and to generate result sets that maximize novelty and/or diversity (Zhai et al., 2003; Agrawal et al., 2009). Unfortunately, unlike the Robertsonian model–whose complexity without indices is $O(|D| \log n)$–for virtually all measures of interest maximizing novelty and diversity is NP-complete (Santini, 2011), so approximate solutions have to be used. No formal, workable definition of novelty and diversity has hitherto been proposed for multimedia.

As a final epistemological note, we point out that while novelty and diversity are often maximized at the same time, they have quite different implications, and affect the results in quite different ways. From the point of view of the final user, novelty should always be maximized, as it avoids receiving redundant results, and uses the "result budget" (the limited number of items that can appear in the result list) to cover different aspects of interest to the user. Diversity is, from the point of view of the user, a nuisance. Each user would of course like to minimize diversity by receiving results only about the interpretation that

she is interested in–the user interested in the Manhattan tribe would be elated to receive only results about the tribe. Maximizing diversity is, on the other hand, in the interest of the server, since the server doesn't know which of the various interpretations each user is interested in, and can only provide data based on global estimates. One can therefore imagine that, with more information about the user, the need for diversity will decrease, while the need for novelty would in any case remain high.

2 MEASUREMENTS OF NOVELTY AND DIVERSITY

While the conceptual definition of novelty and diversity is quite clear, its translation to a precise mathematical formulation has been thus far much more problematic. Most information retrieval work on the subject adopts an operative point of view: a measure function is defined (based on a suitable model of relevance) that fits as well as possible our conceptual understanding of novelty/diversity, and novelty/diversity are, *ex hypothesi*, whatever the function measures. Many of the functions proposed in the literature do measure some form of novelty/diversity, but it is not too clear what combination of the two is being measured. We shall present these measures trying, in the limits in which this is possible, to clarify their relation with the conceptual definitions. We shall also introduce a measure (FZ) in which the two concepts are independently defined, measured, and combined.

We shall use two different models to interpret the relevance of an item for a query. Consider a query q and an item d with relevance $r(d|q) \in [0,1]$. A common interpretation of r is probabilistic: $r(d|q)$ is the probability that a person will consider item d as relevant for query q. Many of the operators in use today are based on this interpretation. In some cases, however, an interpretation based on *degree of truth* is epistemologically more adequate (Dubois and Prade, 2001), an interpretation that requires the formal machinery of fuzzy logic.

As a control group, we shall use two standard information retrieval measures that do not take novelty and diversity into account[1]. Note that a diverse result set will in general score worse than a non-diverse one in these measures. This is to be expected, as from the point of view of the classic measures the best possible result set would be composed of repetitions of the image with the highest relevance. The use of these measures will give us a sense of how much precision are we losing in order to achieve diversity. We assume that the results form a list R of items with relevances $[r_1,\dots,r_n]$, $r_i \in [0,1]$.

Our first non-diversity measure is based on a simple user model. Assume that a user analyzes the list one element at the time, and that, once he reaches position k, she will move on to $k+1$ with probability β, while she will abandon the analysis with probability $1-\beta$. We can then weight each position of the list with the probability that the user will look at it, thus obtaining the *rank based precision*:

$$\mathrm{RBP}(R,k) = \frac{1-\beta}{1-\beta^k}\sum_{i=1}^{k}\beta^{i-1}r_i \tag{1}$$

$\mathrm{RBP}(R,k)$ is the average relevance found by a user that analyzed the list. for our tests, in order to choose β we consider the average number of items seen by a user ($\beta/(1-\beta)^2$). Setting this value to 10 (a reasonable value), we obtain $\beta \approx 0.73$, a value that we shall use throughout the paper. Finally, the *average precision* is defined as

$$\mathrm{AP}(R,k) = \frac{\sum_{j=1}^{k}\left[\sum_{i=1}^{j}\frac{r_i}{j}\right]r_i}{\sum_{i=1}^{k} r_i} \tag{2}$$

None of these measures takes into account diversity; they will form our comparison baseline.

We consider two measures that take diversity into account[2]. The first is derived from the work of (Clarke et al., 2009), modified for our purposes. Assume that an item is characterized by the presence of certain *nuggets* of information, n_μ. Let $w_{k\mu} = P[n_\mu \in d_k]$ the probability of finding nugget n_μ in item d_k. Also, let ω_q be the event "item in position q is found interesting". The probability that ω_q occur due to the fact that d_q contains n_μ is equal to the probability that $n_\mu \in d_q$ that that n_μ is new, that is, that it hasn't been observed in any of the previous items:

$$\begin{aligned} &P[\omega_q | n_\mu \in d_q, d_1,\dots,d_{q-1}] \\ &\quad = P[n_\mu \in d_q]\prod_{j=1}^{q-1}(1-P[n_\mu \in d_j] \\ &\quad = w_{q\mu}\prod_{j=1}^{q-1}(1-w_{j\mu}) \end{aligned} \tag{3}$$

The probability that item d_q will be considered novel is then

$$P[\omega_q | d_1,\dots,d_{q-1}] = \sum_\mu w_{q\mu}\prod_{j=1}^{q-1}(1-w_{j\mu}) \tag{4}$$

[1] During the preparation of this work, we have considered more measures than the ones presented here. For the purposes of this paper, we have retained only those measures that showed statistically significant differences between methods.

[2] See note 1.

Based on the same user model as before, the perceived novelty of a set is

$$\mathrm{NE}(R,k) = \sum_{q=1}^{k} \beta^{q-1} \sum_{\mu} w_{q\mu} \prod_{j=1}^{q-1} (1 - w_{j\mu}) \quad (5)$$

Note that this measures is not significant for $k = 1$, as any set with only one element is novel according to it. For ease of interpretation of the results, we shall normalize it so that is assumes value 0 for $k = 1$, and use the normalized version

$$\mathrm{NNE}(R,k) = \frac{\mathrm{NE}(R,k)}{\mathrm{NE}(R,1)} - 1 \quad (6)$$

NNE that is a pure novelty measure: it determines the non-redundancy of the result set, without taking into account diversity (viz. the way in which the nuggets answer the various interpretations of the query). The measure doesn't attempt to determine whether the result set answers the query or not; typically it is used in conjunction with one of the standard measures to determine at the same time precision and novelty.

The final measure, which we introduce in this paper, is based on a fuzzy model of relevance, in which the relevance $r(d|\tau) \in [0,1]$ is interpreted as the degree of truth of the statement "item d is relevant for topic τ". The idea is to use a BL-algebra (Hájek, 1996) to express two statements about the sets of results. The first states that a set is *diverse* if it covers all topics:

$$\mathcal{D}(R) \equiv \forall\tau.\exists d.r(d|\tau), \quad (7)$$

the second states that a set is *novel* if for each item in it there is a topic that only that item covers (this guarantees that the item is not redundant):

$$\mathcal{N}(R) \equiv \forall d.\exists\tau.(r(d|\tau) \wedge \forall d'.(r(d'|\tau) \rightarrow d = d')) \quad (8)$$

Translating the expression $\mathcal{D} \wedge \mathcal{N}$ into a suitable BL-algebra, one obtains the measure

$$\begin{aligned} \mathrm{FZ}(R,k) \equiv & \bigwedge_{\tau=1}^{T} \bigvee_{d=1}^{k} r(d|\tau) \\ & \wedge \bigwedge_{d-1}^{k} \bigvee_{\tau=1}^{T} \left[r(d|\tau) \wedge \bigwedge_{d' \neq d} \neg r(d'|\tau) \right] \end{aligned} \quad (9)$$

where $a \wedge b = \min\{a,b\}$, $a \vee b = \max\{a,b\}$, and $\neg a = 1 - a$, $a,b \in [0,1]$. Note that, unlike the previous measures, this one does consider both diversity and novelty explicitly, and gives a formal operative definition of the two concepts.

3 FINDING DIVERSE RESULTS

The previous measures give us a way to compare result sets or to obtain an indication of the "goodness" of a result set, but they do not tell us how to find an optimal set. It is however possible to use these measures as the objective function of an optimization algorithm, thereby transforming them from evaluation tools to active result-generating tools. Unfortunately, finding a set that maximizes any of these measures is NP-complete (Santini, 2011) so, in order to manage large data bases, we shall have to resort to an approximation. We use the simplest approximation possible: a greedy algorithm. Given a query q and an item set R, let $s(d|q,R)$ be a function that gives a score for d based on the similarity with q and novelty with respect to R. Suppose that at a certain moment we have collected a partial set of results $R_{k-1} = [d_1, \ldots, d_{k-1}]$. We find our kth result by maximizing $s(d|q,R_{k-1})$ on the rest of the data base, that is, on $D \backslash R_{k-1}$. We add the item that maximizes $s(d|q,R_{k-1})$ to R_{k-1}, obtaining in this way a set R_k, and repeat the process. The algorithm is shown in figure 3

```
diverse(D,q,k)
    R ← ∅;
    for i=1 to k do
      mx ← arg max_{d∈D\R} s(d|q,R);
      R ← R++[mx];
    od
    return r;
```

Figure 3: Greedy algorithm to maximize (approximately) the diversity of a result set. At each time, we add the element of the data base that maximize the diversity with the elements already in the set.

The complexity of this algorithm, without indices on the data base, is $O(k|D|)$.

The algorithm depends on the function $s(d|q,R)$ and in this paper we shall experiment with three such functions, based on three different models: probabilistic, fuzzy, and geometric.

An item shall be described by a vector $w_{k\mu}$, where k is the index of the item and, for each k, $w_k \in \mathbb{R}^T$. Each vector will be normalized, that is, $\sum_\mu w_{k\mu}^2 = 1$. The nature of the coefficients w_μ will vary depending on the data base in use, as we shall see in the following. A query will be expressed by a similar set of coefficients. *A priori* (without diversity) item similarity will be measured using the inner product

$$\mathfrak{q}(d_k,d_h) = \sum_{\mu} w_{k\mu} w_{h\mu} \quad (10)$$

and so will the *a priori* similarity between an item and the query.

In the probabilistic model, we interpret $\mathfrak{q}(d,d')$ as the probability that d and d' be about the same topic. Given a query and a set R of items, the score of an item d will equal the probability that the item is about the same topic as the query and that, at the same time, no item in R is about the same topic as d. This is tantamount to defining

$$s_p(d|q,R) = \mathfrak{q}(q,d) \prod_{d' \in R} (1 - \mathfrak{q}(d,d')) \tag{11}$$

In the fuzzy model we interpret $\mathfrak{q}(d,d')$ as the truth value of the statement *d and d′ are about the same thing.* The function $s_f(d|q,R)$ is then the truth value of the statement *there is a topic in the query for which d is relevant, and no item in R is relevant for that topic*, that is:

$$\exists \tau.(r(d|\tau) \wedge r(q|\tau) \wedge \forall d'(d \neq d' \rightarrow \neg r(d'|\tau)) \tag{12}$$

which translates into the scoring function

$$s_f(d|q,R) = \max_{\tau} \Big[\mathfrak{q}(d,\tau) \wedge \mathfrak{q}(q,\tau) \wedge \min_{d' \in R; d' \neq d} (1 - \mathfrak{q}(d',\tau)) \Big] \tag{13}$$

where $\mathfrak{q}(d,\tau)$ is computed based on a "dummy" item that has the coefficient corresponding to τ set to one and all the others set to 0.

The final model is geometrical. In this model, we consider the query q as a point that endows the space $\mathbb{R}^T$ with a similarity field $\phi(x) = \mathfrak{q}(x,q)$. Our purpose is to fill this space starting with points close to the query q (the point of the data base in which $\phi(x)$ is maximum) but without choosing points too close to one another, compatibly with the necessity of staying similar to the query. In order to do this, we fill the space with *similarity holes*. Each item $d \in R$ will generate a dissimilarity field around it that will reduce the field ϕ in its vicinity. Assuming that this field is Gaussian, the similarity function that we use is

$$s_g(d|q,R) = \mathfrak{q}(d,q) \prod_{d' \in R} (\alpha + (1-\alpha) \exp\left[-\frac{\mathfrak{q}(d,d')^2}{2\sigma^2} \right]) \tag{14}$$

where α is a small "residual" value that avoids that the similarity field be zero in correspondence of a previous result. Typically, $\alpha \in [0, 0.1]$.

4 TESTS

In order to evaluate our diversity methods using the given measures, we have to begin with two methodological choices: (1) whether to conduct a user study and use formal measures on the result set, and (2) whether to use a large uncontrolled data base (such as that provided by internet search services) or a smaller, controlled one. A user study is clearly not appropriate in this case: a person responds always to a whole system and to to a measure or an algorithm; embedding our algorithm into a system would create a number of extra variables too large to control.

As to the large uncontrolled data sets, they would lead to poor experimental design: a good design must allow the experimenter to impose a treatment on a group of objects while controlling the statistical variables that are not being measured. This would be impossible in a large web-based data base, therefore such a measurement would qualify as an observation but not as an experiment, yielding at best anecdotal evidence.

We check the three diversity alternatives against Robertsonian retrieval in two scenarios: the Im2Text data set (Ordoñez et al., 2011), which contains 1.000.000 annotated images (in which retrieval is done based on the annotation text) and the Event data set (Li and Fei-Fei, 2007). In this case, retrieval is based on visual information using the features of (Ciocca et al., 2012).

4.1 Annotated Data Base

The first test is carried out on a data base composed of 1,000,000 annotated images taken from the web site *flickr*(TM) (Ordoñez et al., 2011). Each image is associated with a short text (1 to 12 words). The people who wrote the texts were under no obligation to describe the contents of the relative image, although in the majority of cases the text contains clues to the contents of the image.

The text of each image was processed removing the stop-words and stemmed so as to obtain a collection of stems for each image. At the end of this phase, all images had at least one stem left, so it was not necessary to prune the data set. Weighting was done using the standard tf-idf scheme: if stem μ appeared n_j times in image d_j and it appeared in N_μ images of the collection, its weight was, for image d_j, $w_{j\mu} = n_j / \log N_\mu$. Finally, the weights vector of each image was normalized so that $\|w_j\| = 1$. The query, consisting of a set of keywords, was similarly processed. Image similarity is given by the inner product. The Robertsonian results were obtained sorting with respect to this similarity, while for the diversity evaluation, we used the diverse similarity $s(d|q,R)$ and the "diverse" algorithm. In the measures that need a separation in topics, we made the approximation that each word represented a separate topic. This is an approx-

Table 1: Results, on the four measures, for diversity retrieval on the annotation data base for result sets of $\{5, 10, 15, 20\}$ images and for the Robertsonian, probabilistic, fuzzy, and geometric retrieval models. The measures are computed on a set of eight single-word queries representing simple objects and concepts (car, house, friend, sea, person, tree, clock, dress).

	k	Robert.	probab.	fuzzy	geom.
RBP	5	0.8	0.34	0.53	0.52
	10	0.76	0.24	0.54	0.47
	15	0.76	0.19	0.53	0.44
	20	0.76	0.17	0.53	0.42
AP	5	0.85	0.72	0.72	0.74]
	10	0.83	0.63	0.65	0.67]
	15	0.82	0.57	0.62	0.63]
	20	0.81	0.53	0.61	0.6
NE	5	0.31	0.77	0.7	0.71
	10	0.33	1.0	0.94	0.97
	15	0.33	1.05	0.99	1.03
	20	0.33	1.06	1.0	1.04
FZ	5	0.09	0.38	0.4	0.43
	10	0.09	0.25	0.29	0.41
	15	0.09	0.2	0.29	0.40
	20	0.06	0.17	0.22	0.38

imation, as there may be synonyms or words related to the same concept. We haven't measured the effect of this simplifying hypothesis, but it is likely that it will not affect the result too much: the vocabulary used in the annotations is fairly poor (some 100,000 different terms form a total of about 5,000,000 overall words) and uniform, not liable to present massive polysemy. We have used a randomized experimental design: for every measure, we have chosen 100.000 images at random from the data set, and we have repeated the experiments ten times with different randomized samples.

For the first test, we presented eight single-word queries representing common objects of the type likely to appear in this type of data set: *car, clock, dress, friend, house, person, sea*, and *tree*. The results were analyzed using teh four measures presented in section 2, using ANOVA ($p = 0.01$) to determine significance, and are shown in table 1. The two non-diversity measures (RBP, AP) reveal that, as expected, the introduction of diversity reduces the precision of the results. The highest drop (up to 78%) is that of the probabilistic method with RBP; the fuzzy method is always below a 30% loss, while the geometric reaches a 45% loss in the RBP with $k = 20$. The other measures (NE, FZ) show, more predictably, an increase when applied to diversity-enforcing methods. The probabilistic model performs especially well in the NE measure, while it seems to perform worse than the other models in FZ. This should not be surprising, as NE is based on a probabilistic model, while FZ measures the performance in terms of fuzzy logic using the same model as the fuzzy similarity measure.

As a curiosity, we show in figure 4 the result of the query *car* on the same data base as in figure 2 using the fuzzy model of diversity. The results contain a wider variety of examples (because of the way it works, the first result of the diversity algorithm is always the same as that of the Robertsonian ranking), including a *car port* (8th image) and a car picture. The somewhat enigmatic second result is an image that the author had labeled *4 possible designs I will put into my seats and/or somewhere in my car*. After stop-word removal and stemming, the only remaining keywords for this example were *car, design, seat*. We should remark that examples like this are only anecdotal curiosities and have no scientific validity. They shouldn't be taken too seriously.

Figure 4: A result corresponding to the query *car* on an annotated data base using the fuzzy diversity algorithm.

As we have mentioned in section 2, diversity measures don't take into account the quality of the results vis-à-vis the query, and they should be considered in conjunction with standard quality metrics. We do this in figure 5; for each one of the four models (Robertsonian, probabilistic, fuzzy, geometric) we derived two diagrams, plotting a precision measure one versus a diversity one.

Figure 5(a) shows NE plotted against RPB. The two were plotted together as they are both based on the same user model. Figure 5(b) shows AP verus FZ. Robertsonian retrieval behaves quite as expected: the non-diversity measure has a high value and the diversity measure has a low one. Neither one changes much with k. The probabilistic and geometric models show, as k increases, an increase in diversity and a decrease in precision, while the value of RBP for the fuzzy model remains constant. Notice that while NE increases as k increases, FZ decreases. This behavior derives from the different aspects of novelty that the two measures focus on. NE determines the probability that a user interest in any of the available topics will find something useful in the results, a probability

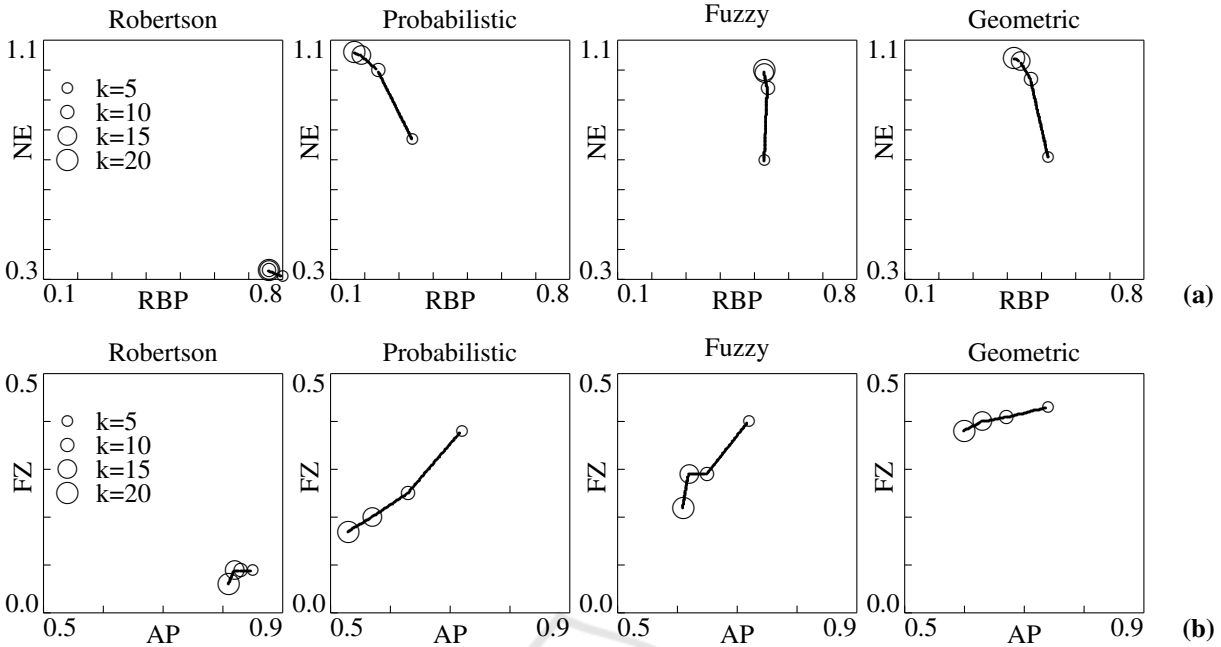

Figure 5: Plots of one diversity measure vs. a precision one for queries consisting of words defining concrete, specific objects. In (a), NE is plotted against RBP for the four models (Robertsonian, probabilistic, fuzzy, geometric), in (b) FZ is plotted against AP. The curves represent increasing values of k, from $k = 5$ to $k = 20$. For an analysis of the curves, see the text.

that increases with the number of results; FZ determines whether an item is supplying something new to the query, something that no other items of the set have supplied. As the topics of interest are being covered, redundancy grows, so smaller sets will have in general higher scores than larger ones (this is not a problem for the use of the measure as its purpose is to compare sets of the same size). The drop in the non-diversity measure (AP) is much more pronounced for the probabilistic model than for the fuzzy and geometric.

The queries of figure 5 consisted of concrete objects, stuff that can be found in the images (or, more cogently, in their descriptions). A further series of tests was carried out using, as queries, abstract concepts such as*freedom*, *sadness*, or *joy* (figure 6). In this case, the performance of Robertson is quite different from that of figure 5: it does indeed provide some diversity and its precision (both RBP and AP) decreases when k increases. The qualitative results for the probabilistic, fuzzy, and geometric models are similar to those of figure 5, the main difference being the smaller range of change of the diversity measures. Note that on the FZ measure the Robertsonisn model scores better than the probabilistic one, that is, the probabilistic model introduces more redundancy than the Robertsonian; it is a behavior that we shall find again in visual queries.

4.2 Visual Queries

Our second set of tests takes on the problem of retrieval based on visual similarity, without annotation. We are using the Event data set of (Li and Fei-Fei, 2007), originally developed for testing event classification systems. On this data set, we do content based retrieval using the feature vector presented in (Ciocca et al., 2012). These features are similar in principle to other systems based on the output of suitably trained classifiers, such as *classemes* (Torresani et al., 2010) or Li *et al.*'s *Object Bank* (Li et al., 2010).

The feature vector consists of the output of 56 classifiers, representing 14 different classes and four different low-level features. If $i = 1,\ldots,4$ are the four low-level features and $j = 1,\ldots,14$ the classes, then the feature vector is represented as $\phi = [\phi_{ij}]$, where ϕ_{ij} is the output of a classifier that receives as input the ith feature and is trained to recognize the jth class. The four low-level features are a block color histogram, a global histogram, an edge direction histogram, and a bag-of-word representation obtained using SIFT (we have two global and two local features, two color and two shape features); the classifiers are support vector machines with Gaussian kernels, and the 14 categories are a varied set consisting mostly of scene-based categories (animals, city, close-up, desert, flowers, forest, indoor, mountain, night, people, rural, sea, street, and sunset). The outputs of the 56 classifiers form a

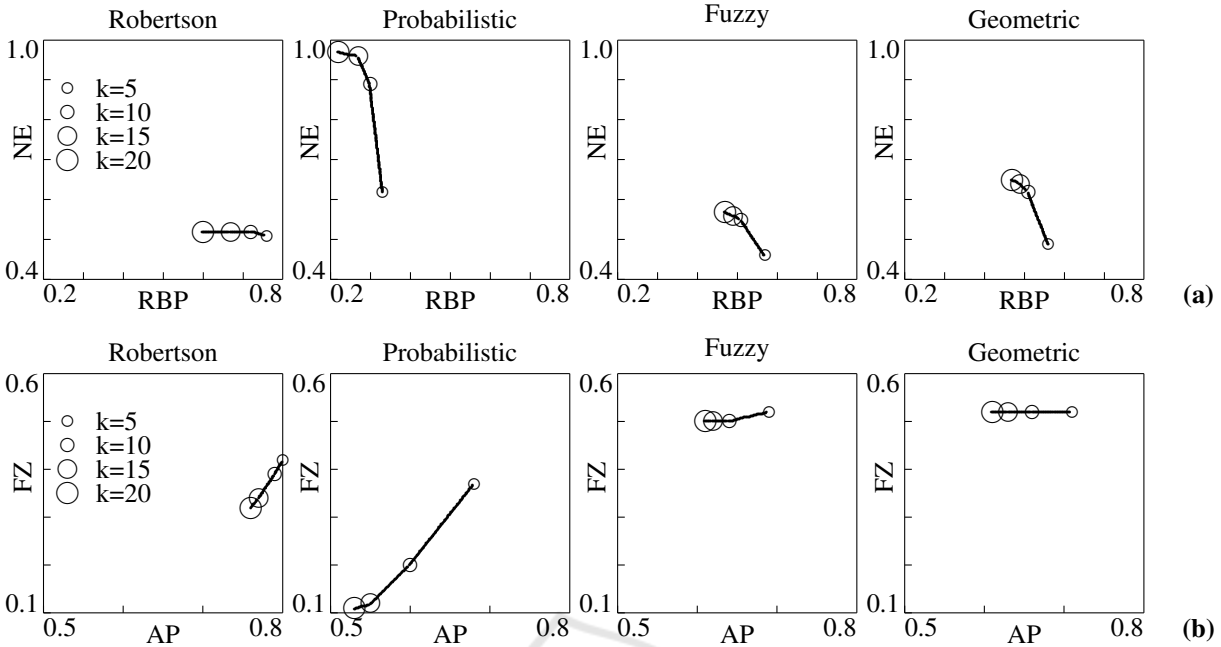

Figure 6: Plots of one diversity measure vs. a precision one for queries consisting of words defining abstract concepts. In (a), NE is plotted against RBP for the four models (Robertsonian, probabilistic, fuzzy, geometric), in (b) FZ is plotted against AP. The curves represent increasing values of k, from $k = 5$ to $k = 20$. For an analysis of the curves, see the text.

56-dimensional *prosemantic* feature space, which we use as a basis of distance-based retrieval using a Euclidean distance.

For the purpose of diversity, we consider each category as a concept, so we must somehow group together the output of the four classifiers corresponding to each one of them. That is, from the output of the four classifiers $\phi_{1,1} \ldots, \phi_{4,j}$ we derive an indicator of the presence of concept j: $\psi_j = f(\phi_{1,1} \ldots, \phi_{4,j})$. In order to derive the indicators ψ_j we use the same two interpretations that we have used in the previous section: probabilistic and fuzzy. In the probabilistic interpretation, the probability that the jth category be represented in the image is equal to the probability that at least one of the four classifiers associated to the category detect it, that is:

$$\psi_j^{(p)} = 1 - \prod_{i=1}^{4} (1 - \phi_{i,j}) \tag{15}$$

while in the fuzzy interpretation, the truth value of the statement "the image belongs to category j" is the disjunction of the statements corresponding to the four features:

$$\psi_j^{(f)} = \max_{i=1,4} \phi_{i,j} \tag{16}$$

These coefficients are interpreted as concept weights, and used to determine similarity and topic relevance exactly as in the case of the annotation data base.

The query, in this case, consisted in one of the images of the data base (query by example). The results are shown in figure 7. The behavior of NE is almost the same for all models. In this case, the inherent imprecision of visual retrieval increments diversity even without special provisions for doing so (remember that NE measures the probability that the user will find something interesting without considering redundancy, so imprecision is good for NE). In the case of FZ, which does measure redundancy, the Robertsonian model performs worse than fuzzy and (partially) geometric, while it performs better than probabilistic. The results seem to indicate that the inherent imprecision of visual retrieval does introduce a good degree of diversity even in the absence of specific diversity-enforcing methods, but it does so at the expense of a great redundancy. The introduction of specific means to enforce novelty allows us to obtain the same diversity without redundancy. The reason why the fuzzy model works better than the probabilistic might indeed be related to redundancy: the fuzzy model tries explicitly to enforce novelty and, therefore, explicitly reduces redundancy, while the probabilistic model is more focused on diversity.

5 CONCLUSIONS

The concepts of *novelty* and *diversity* were introduced in information retrieval as a tool to make result sets more informative by covering the different interpreta-

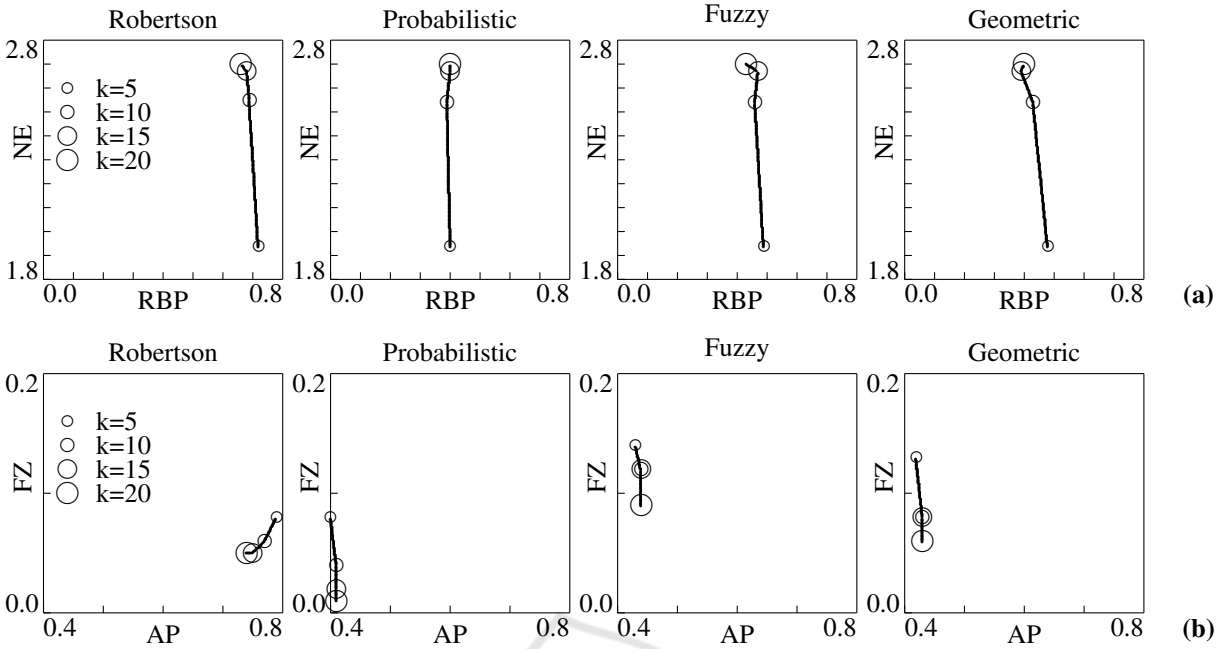

Figure 7: Plots of one diversity measure vs. a precision one for viosual queries. In (a), NE is plotted against RBP for the four models (Robertsonian, probabilistic, fuzzy, geometric), in (b) FZ is plotted against AP. The curves represent increasing values of k, from $k = 5$ to $k = 20$. For an analysis of the curves, see the text.

tions resulting from the ambiguity of a query and the different aspects resulting from its underspecification. In multimedia retrieval, these concepts can be useful in order to avoid scarcely informative result sets that may be a consequence of the presence of semantically similar images.

In this paper, we have adapted the ideas of novelty and diversity to the specific needs of multimedia information. We have given an operative definition in the form of a number of measures, and we have defined a general algorithmic schema for finding diverse and novel result sets. We have instantiated this schema using three models: a probabilistic one, a fuzzy one, and a geometric one, and we have conducted a series of tests to determine their behavior.

The results indicate that novelty and diversity are very useful concepts to use especially in the case of tag- or annotation-based repository. In the case of visual query the inherent imprecision of the methods provides varied results even in the absence of specific provisions, but in the absence of specific novelty-enforcing methods, this comes at the expense of a considerable redundancy. Since diversity is "enforced" by the imprecision of the search, models that try explicitly to increase novelty, such as the fuzzy model, work better for visual retrieval than methods that work on an undifferentiated mix of novelty and diversity, such as the probabilistic.

REFERENCES

Agrawal, R., Gollapudi, S., Halverson, A., and Leong, S. (2009). Diversifying search results. In *Proceedings of WDSM '09*. ACM.

Chapelle, O., Metzler, D., Zhang, Y., and Grinspan, P. (2009). Expected reciprocal rank for graded relevance. In *Proceedings of the 18th International Conference on Information and Knowledge Management*. ACM.

Ciocca, G., Cusano, C., Santini, S., and Schettini, R. (2012). Prosemantic image retrieval. In *European Conference on Computer Vision*, pages 643–6. Springer.

Clarke, C., Kolla, M., and Vechtomova, O. (2009). An effectiveness measure for ambiguous and underspecified queries. In *Proceedings of ICTIR 2009*, Lecture Notes in Computer Science. SPringer-Verlag.

Dubois, D. and Prade, H. (2001). Possibility theory, probability theory and multiple-valued logics: A clarification. *Annals of Mathematics and Artificial Intelligence*, 32(1-4):35–66.

Hájek, P. (1996). Basic fuzzy logic and BL-algebras. Technical Report V736, Institute of Computer Science, Academy of Science of the Czech Republic.

Li, L. and Fei-Fei, L. (2007). What, where and who? classifying events by scene and object recognition. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 1–8.

Li, L., Su, H., Xing, E., and Fei-Fei, L. (2010). Object bank: A high-level image representation for scene classification and semantic feature sparsification. *Advances in Neural Information Processing Systems*.

Ordoñez, V., Kulkarni, G., and Berg, T. L. (2011). Im2text: Describing images using 1 million captioned photographs. In *Neural Information Processing Systems*.

Rasiwasia, N., Jose, C. P., Coviello, E., Doyle, G., Lanckriet, G. R. G., Levy, R., and Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. In *Proceedings of ACM Multimedia*. ACM.

Robertson, S. E. and Spark-Jones, K. (1976). Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27:129–46.

Santini, S. (2011). Efficient computation of queries on feature streams. *ACM Trans. on multimedia computing, communications and applications*, 7(4).

Santini, S. and Castells, P. (2011). An evaluation of novelty and diversity based on fuzzy logic. In *International Workshop on Novelty and Diversity in recommender systems (part of RecSys 2011)*.

Saracevic, T. (2007). Relevance: a review of the literature and a framework for thinking on the notion of information science. *Journal of the American Sociery of Information Science and Technology*, 58(13):2126–44.

Torresani, L., Szummer, M., and Fitzgibbon, A. (2010). Efficient object category recognition using classemes. In *Proceedings of the European Conference on Computer Vision*, number 6311 in Lecture Notes on Computer Science, pages 776–89. Springer-Verlag.

Zhai, C. X., Cohen, W. W., and Lafferty, J. (2003). Beyond independent relevance: methods and evaluation metric for subtopic retrieval. In *Proceedings of the 26th International ACM SIGIR Conference in Research and Developmens in Information Retrieval*, pages 10–7. ACM.