A study on the evaluation of relevance feedback in multi-tagged image datasets

Roberto Tronci
Amilab - Laboratorio Intelligenza d'Ambiente
Sardegna DistrICT, Pula, Italy
roberto.tronci@sardegnaricerche.it

Luisa Falqui, Luca Piras, Giorgio Giacinto
DIEE - University of Cagliari
Piazza d'Armi, Cagliari, Italy
luca.piras@diee.unica.it, giacinto@diee.unica.it

***Abstract*—This paper proposes a study on the evaluation of relevance feedback approaches when a multi-tagged dataset is available. The aim of this study is to verify how the relevance feedback works in a real-word scenario, i.e. by taking into account the multiple concepts represented by the query image. To this end, we first assessed how relevance feedback mechanisms adapt the search when the same image is used for retrieving different concepts. Then, we investigated the scenarios in which the same image is used for retrieving multiple concepts. The experimental results shows that relevance feedback can effectively focus the search according to the user's feedback even if the query image provides a rough example of the target concept. We also propose two performance measures aimed at comparing the accuracy of retrieval results when the same image is used as a prototype for a number of different concepts.**

***Keywords*-content based image retrieval; relevance feedback; correlation measures.**

I. Introduction

The task of retrieving images from multimedia collections is one of the most active topics in computer science nowadays. In many cases, e.g., a raw collection of images from a photo repository, the only information that can be extracted is the related visual content. In these cases the retrieval task must rely only on low-level features, and high-level categories (e.g., indoor, faces, sport, etc.). As a consequence, the image retrieval task becomes more difficult than in these cases due to the semantic gap between the visual content (i.e., the data used for the retrieval process) and the semantic concepts (i.e., the goal of the retrieval process).

For these reasons, different Content Based Image Retrieval techniques (CBIR) had been proposed in the literature to improve the retrieval performances [1], [2], [3]. The effectiveness of a CBIR technique strongly depends on the choice of the set of visual features, and on the choice of the metric used to model the users' perception of image similarity. Unfortunately, even if we extract the "best" visual features for a given retrieval problem, or we design a good similarity metric, the set of retrieved images often fits the users needs only partly. To overcame this problem, the use of Relevance Feedback has been widely studied as a tool to allow users refining the results by submitting a feedback on the images' relevance [4], [5], [6], [7].

Relevance Feedback (RF) is a mechanism that involves directly the user by allowing her to refine the retrieval results by marking the retrieved images for a given query as relevant or non-relevant. Then, this feedback is exploited to "adjust" the retrieval mechanism, and it is used to propose to the user a new set of images that is deemed to be relevant according to the given feedback. Typically, the performance of RF mechanisms is evaluated by means of single-tagged datasets. With the term single-tagged we mean that each image of the dataset is associated to just one tag (i.e. the label associated to a visual concept) even if the image contains more than one visual concept. It is the opinion of the authors that the assessment of the capabilities of RF techniques carried out by this kind of datasets are somehow limited. This limitation is related to the fact that only a single tag/concept is associated to each image in the dataset. Commonly, this "limitation" is accepted for two reasons: first, the main visual concept of the image is usually considered as it is easier to find; second, it is more easy, during the testing phase, to simulate the behavior of users that submit the feedback with respect to only one concept. Despite the fact that these reasons allow to set up a simple evaluation scenario, they don't allow to full evaluate the potential behind RF techniques because an image is typically associated to a number of concepts, and because this modality of simulated feedback is not a good model of the behavior of a real user.

Unfortunately, the simulation of a real user's feedback in a real case scenario (i.e. when different visual concepts are associated to a single image) is quite difficult. Following the query-by-example paradigm, the user submits to the CBIR system an image as query, and the system retrieves the images that are most similar to the query from the visual point of view. When the user is asked to mark the images as being relevant or not, then a real user can be interested in retrieving images related to more than just one of the visual concepts represented in the query image. Thus, different retrieval tasks can be performed starting from a given query image, as many as the possible combinations of visual concepts contained in the image itself. This kind of behavior is quite difficult to simulate unless you have a toy dataset, or a large number of real users agree to perform a live experimentation.

In this paper we address the above topic, and propose

some way to simulate some of the possible behaviors of a real user in different scenarios, i.e., the logic employed by the user to mark the images as relevant or not. In our opinion these scenarios can be used to a more thorough assessment of RF techniques. In order to perform such an assessment, we also propose some novel concept correlation measures, aimed at assessing if some different concepts that typically appear in different images, can be retrieved by relying on just one of the concepts. In fact, while in the case of a single-tagged dataset the retrieval process is clearly driven by just one concept, in the case of a multi-tagged dataset it can be of interest to assess if the search for multiple concepts is actually driven by a few of them, or if these multiple concepts actually represent a higher level concept.

The rest of the paper is organized as follows: in Sec. II we propose new scenarios for testing RF approaches; in Sec. III we propose some novel performance measures to be used in the proposed scenarios; in Sec. IV the RF techniques used in this paper are briefly reviewed, while in Sec. V all the details of the experimental phase are given.

II. A "NEW" MODUS OPERANDI OF TESTING A RELEVANCE FEEDBACK METHODOLOGY

In the field of CBIR research, the use of real world (i.e. multi-tagged) datasets is progressively increasing because they allow for a more "real user-like" evaluation of the algorithms developed for CBIR. The use of the multi-tagged image datasets have some drawbacks, especially in the case of the assessment of relevance feedback methodologies, as they complicate the simulation of a user's feedback. As stated in the introduction, the usual scenario for a relevance feedback experiment is to take into account a single-tagged image dataset. In this scenario, it is easy to automatically "simulate" the feedback given by user, as the query image is associated to just one concept, and the simulated user marks as relevant all the images that belong to the same concept of the query, and marks as non-relevant all the other ones. On the other hand, in a real-world scenario, each image is associated to more than one concept. In this case, can we simulate the behavior of a user giving her feedback? If yes, how? If no, are there some ways to partially simulate it? Let us try to answer these questions with an example.

Let us take into account a dataset where each image is associated to different concepts $L = \{l_1, l_2, \dots l_n\}$. Let us take a query image q from the dataset that has associated the concepts $\{l_1, l_2, \dots l_k\}$, where $k < n$. Users who starts the search using q as a query can follow one of the four scenarios described below:

Single concept scenario: this scenario is related to users that are interested to just one of the k concepts $\{l_1, l_2 \dots l_k\}$. This scenario is quite similar to the usual scenario, with the difference that each image belongs to different concepts, thus different set of images can be relevant to the query depending on the selected concept.

Multiple concept AND scenario: in this scenario users are interested in images that exhibits an exclusive combination of the concepts related to the query (pairs, terns, etc.) $\{l_1 \wedge l_2, \dots l_{k-1} \wedge l_k, l_1 \wedge l_2 \wedge l_3, \dots\}$. In this scenario the search for similar images is driven by the combination of concepts. It is easy to see that the search for multiple concepts is analogous to the search of a single higher level concept made up of the combination of concepts l_c.

Multiple concept OR scenario: in this scenario users are interested in images that exhibits a non-exclusive combination of the concepts related to the query $\{l_1 \vee l_2, l_1 \vee l_3, \dots l_2 \vee l_3 \vee l_k, \dots\}$. Thus, an image is relevant if it exhibits at least one of the concepts the user is looking for. E.g., in the case of $\{l_1 \vee l_2 \vee l_k\}$, at each interaction the images that are relevant to the user's query are those who belong to one of the concepts $\{l_1, l_2, l_k\}$ or to a combination of them.

Multiple concept AND-OR scenario: this is the combination of the previous two scenarios. In this case the searching target is a combination of concepts connected by AND and OR statements: e.g., a possible target is $\{l_1 \vee (l_2 \wedge l_3)\}$, and the related relevant images are those one that have l_1 or l_2, l_3 or l_1, l_2, l_3 as concepts.

This four scenarios cover all possible behavior of a real user. All of these scenarios assume that the user holds the same behavior for a given query, and for each feedback interaction. In a real case, a user can also switch between scenarios in different interactions, e.g. she starts using the pair $\{l_1 \wedge l_2\}$, while afterward she starts marking as relevant also the images that have only the l_1 concept.

The *single concept scenario* allows testing the ability of a relevance feedback methodology not only to adapt the search to the feedback during the marking interaction, but also to verify the capability of adaption to different target concepts even if the query image is the same.

The *multiple concept AND scenario* tests the ability of a relevance feedback methodology to refine the search exploiting the feedback when the target is a combination of visual concepts. From a formal point of view, the *AND* of different concepts produces a new concept l_c (i.e. is formally equivalent to the previous one).

The *multiple concept OR scenario* and *multiple concept AND-OR scenario* verify the ability of a relevance feedback methodology to refine the search exploiting the feedback when the target is a "complex" combination of visual concepts.

III. MEASURES FOR CORRELATION IN MULTIPLE CONCEPT QUERY SCENARIOS

The scenarios proposed in the previous section can offer different point of views for the analysis of a relevance feedback methodology, and they allow to use real image datasets where each image is associated to more than one concept. All the performance measures developed so far

(e.g. precision, recall, F_1, etc.) are good for estimating the performance also in these scenarios.

In the case of the multiple concept scenarios, when we are dealing with a combination of concepts, some questions arise. As partially mentioned in the previous sections, RF approaches work in this way: the system present to the user a limited number of images n; then the user marks these images as relevant or not, and this feedback is exploited by the RF method to recalibrate the retrieval result and propose "new" images to the user (the user can iterate the process as many times she wants). For a multi-tagged dataset, the feedback is usually provided by means of a combination of concepts: in this case, is it the operation of recalibrating the results driven by the combination of concepts or by a subset of them from a formal point of view? Moreover, when we are looking for a combination of concepts, is the feedback given for a combination of concepts better than the feedback given for the individual concepts in terms of relevant images that are found after the recalibration step? These questions doesn't regard the performance of the system itself, but they are useful to analyze in a deeper way the behavior of the system when it involves multiple concepts.

To provide an answer to the previous questions, we propose to use the following *concept correlation* measures, C_1 and C_2. These measures are formulated in the case of two concepts, but they can be easily extended to more concepts.

$$C_1 = \frac{1}{2}\left(R_{l_1}(l_1 \star l_2) + R_{l_2}(l_1 \star l_2)\right) / \left(R_{l_1 \star l_2}(l_1 \star l_2)\right) \quad (1)$$

$$C_2 = \frac{1}{2}\left(R_{l_1}(l_1 \star l_2)/R_{l_1}(l_1) + R_{l_2}(l_1 \star l_2)/R_{l_2}(l_2)\right) \quad (2)$$

where $l_1 \star l_2$ is the type of combination chosen (i.e.. it depends on the scenario we want to test), and $R_x(y)$ is the number of retrieved images that are associated to the concept y, while relevance feedback is provided according to concept x; i.e. in Equation (1), $R_{l_1}(l_1 \star l_2)$ is the number of images that exhibit the concept $l_1 \star l_2$ among those that are retrieved by relevance feedback when the images that exhibit concept l_1 are marked as relevant.

The *concept correlation* measure C_1, Eq. (1), answers the questions proposed above: i.e., for a given combination of concepts $l_1 \star l_2$, it measures if it is better to have the feedback driven by the combination rather than by the single concepts. Thus, if a large value of C_1 is attained, thus it means that we can retrieve a larger number of relevant images according to the combined concept by marking as relevant just the individual concepts (separately), instead of marking as relevant the combination of them. If the value of C_1 is small, thus it means that the opposite holds.

The *concept correlation* measure C_2, Eq. (2), shows how much a single concept is connected with the combined concept in the relevance feedback retrieval process. For example, it can be used to see how much the combination of the concept is "embedded" in a single concept. In the case of the *multiple concept AND scenario* when the value of C_2 is equal to 1 it means that in the retrieved images have both the tags l_1 and l_2, thus the two tags are highly correlated.

IV. Relevance feedback techniques for this experimental set-up

In this section the relevance feedback techniques used in the experimental phase, are described: one is based on the nearest-neighbor paradigm, the other is based on Support Vector Machines. The use of the nearest-neighbor paradigm is motivated by its use in a number of different pattern recognition fields, where it is difficult to produce a high-level generalization of a class of objects, but where neighborhood information is available [8], [9]. In particular, nearest-neighbor approaches have proven to be effective in outliers detection, and one-class classification tasks [10], [11]. Support Vector Machines are used because they are one of the most popular learning algorithm when dealing with high dimensional spaces as in CBIR [12], [13] .

1) k-NN relevance feedback: for the nearest neighbor we resort to a technique proposed in [7] where a score is assigned to each image of a database according to its distance from the nearest image belonging to the target class, and the distance from the nearest image belonging to a different class. This score is further combined to a score related to the distance of the image from the region of relevant images. The combined score is computed as follows:

$$rel^f(I) = \left(\tfrac{n/k}{1+n/k}\right) \cdot rel_{BQS}(I) + \left(\tfrac{1}{1+n/k}\right) \cdot rel_{NN}(I)$$

where n and k are the number of non-relevant images and the whole number of images retrieved after the latter iteration, respectively. The two terms rel_{NN} and rel_{BQS} are computed as follows:

$$rel_{NN}(I) = \frac{\|I - NN^{nr}(I)\|}{\|I - NN^{r}(I)\| + \|I - NN^{nr}(I)\|}$$

$$rel_{BQS}(I) = \left(1 - e^{1 - \frac{d_{BQS}(I)}{\max_I d_{BQS}(I)}}\right) / (1 - e)$$

where $NN(I)$ denotes the nearest neighbor of I, $\|\cdot\|$ is the metric defined in the feature space at hand, and d_{BQS} is the distance of image I from a reference vector computed according to the Bayes Decision Theory [14]. If we are using F feature spaces, we have different scores $rel(I)$ for each f feature space. Thus the following combination is performed to obtain a "single" score:

$$rel(I) = \sum_{f=1}^{F} w_f \cdot rel^f(I)$$

where the w_f is the weight associated to the f-space.

The weights w_f are estimated by taking into account the minimum distance between all the pairs of relevant images

[15], and the minimum distance between all the pairs of relevant and non-relevant images as follows

$$w_f = \frac{\sum_{i \in R} d^f_{min}(I_i, R)}{\sum_{i \in R} d^f_{min}(I_i, R) + \sum_{i \in R} d^f_{min}(I_i, N)}$$

2) SVM relevance feedback: Support Vector Machines are used to find a decision boundary in each feature space $f \in F$. The use of an SVM for this tasks is very useful because, in the case of image retrieval, we deal with high dimensional feature spaces. For each feature space f, each SVM is trained using the feedback given by the user. The results of the SVMs in terms of distances from the hyperplane of separation are then combined into to a relevance score through the Mean rule as follows

$$rel_{SVM}(I) = \frac{1}{F} \sum_{f=1}^{F} rel^f_{SVM}(I)$$

V. Experimental phase

In these experiments we will show the tests on the following scenarios described in Section II: the *single concept scenario*, the *multiple concept AND scenario*, and the *multiple concept OR scenario*. The aim of these experiments is to take into account a multi-tagged image dataset, and verify how different relevance feedback techniques work in the proposed scenarios. The *multiple concept AND-OR scenario* is not showed for lack of space.

A. Dataset and experiments setup

For the purpose of testing the scenarios proposed in Section II, the MIRFLICKR-25000 collection [16] had been chosen. This collection consists of 25000 images downloaded from the social photography site Flickr through its public API. The average number of tags per image is 8.94. In the collection there are 1386 tags which occur in at least 20 images. Moreover for the images also some manual annotations are available (24 in the collection used). We chose 11 feature spaces to be used: i.e., the *Scalable Color Descriptor*, the *Color Layout Descriptor*, the *Edge Histogram Descriptor*, the *RGB Histogram*, the *HSV Histogram*, the *RGB Histogram*, the *Fuzzy Color and Texture Histogram*, the *Color and Edge Directivity Descriptors*, the *Tamura Texture*, the *JPEG Coefficient Histogram*, the *Gabor Texture*, and the *Appearance-Based Image Features*.

We analyzed all the tags of the collection by a semantical point of view, and fused the tags with the annotations in a tag verification process. After this process, we decided to keep only the tags which occur in at least 100 images. This decision were taken because we aimed to have the single concepts adequately represented in the dataset used in the evaluation experiments. This process of fusing and discarding tags brought us to keep 24718 images and 69 tags, with an average number of tags per image of 4.19. Then, we evaluated all the possible combination of concepts that derive from the modified collection for testing the multiple concept scenarios. The result of the evaluation was that only the pairs of concepts were worth to be used in a experimental phase, i.e., the number of terns, and higher combinations were shared by small subsets of images to be worth of being used in an experimentation. Thus, for the pairs of concepts we kept 658 of them which occur in at least 25 images, for the purpose of testing the *multiple concept AND scenario* and *multiple concept OR scenario*. Finally as query images we chose 1.294 of them from the refined collection. These query images have a number of tags per image that varies from 3 to 10, with an average number of tags per image equal to 4.69 (thus very similar to the value in the all collection).

For each one of the 1.294 query image, a relevance feedback experiment has been performed by using all the tags and pairs associated to the images as a target, i.e., given a query image, we considered, one at a time, each single tag and each pair of tags as target of the retrieval process to be refined through the relevance feedback. In this way we performed over 17.201 relevance feedback experiments per relevance feedback technique. Each experimentation consists of 10 iterations: the first one is based on a nearest neighbor on all the feature spaces, and the other 9 are iterations based on one of the relevance feedback methodology described above. At each iteration $n = 20$ images are "shown to the user" for marking the feedback.

As already said in Section II, for the aim of evaluating the performance in the proposed scenarios all the well-known performance measures can be used. Then, the performance of the experiments will be assessed using the *Precision*, a modified definition of the Recall, that we named "user perceived" *Recall*, and the *concept correlation* measures C_1 and C_2 proposed in Section III. The "user perceived" *Recall* is a recall measure that takes into account just the maximum number of relevant images that can be shown to the user according to the number of iterations, and the number of images displayed per iteration, and it is computed as follows

$$r_p = \frac{A \cap R}{R^*} \quad , \quad R^* = \begin{cases} R & \text{, if } < n \cdot i \\ n \cdot i & \text{, otherwise} \end{cases}$$

where A is the number of images at the iteration i, R is the number of relevant images in the dataset (for a given target), and n is the number of images shown per iteration.

B. Experimental results

In this case of *single concept scenario* we started from each one of the 1.294 query images and tested all the possible concepts as single target. Every single concept belonging to a query image was used to simulate a user that it is looking for images similar to the query according to that concept. Thus, each query image has been used

as a starting example for different retrieval tasks. In this way, 6.070 retrieval tasks were performed for each relevance feedback technique. We compared the relevance feedback techniques described in the previous section, i.e., the k-NN based (NN in the figures) , and the SVM, with *browsing*. The *browsing* is nothing more than the showing the user the n images nearest to the query with no feedback. The aim of comparing relevance feedback with *browsing* is to show the benefits of relevance feedback. To put it simple: can a relevance feedback approach retrieve more relevant images than simply *browsing* the collection by sorting the images according to the visual similarity with the query?

In Table I the average results in terms of *Precision*, and "user perceived" *Recall* are reported for this scenario. The results show that, as the number of iterations increase, the performance of the relevance feedback methods increase, as well as the difference in performance with the *browsing*. From these analysis it turns out that the SVM exhibits the biggest increasing performance power.

Precision

it.	0	1	2	3	4	5	6	7	8	9
SVM	29.6	28.0	**30.5**	**33.1**	**35.2**	**36.9**	**38.3**	**39.4**	**40.4**	**41.1**
NN	29.6	**28.7**	29.4	30.0	30.6	31.1	31.5	31.8	32.1	32.4
browsing	29.6	28.6	28.1	28.0	27.7	27.5	27.3	27.2	27.1	26.9

Recall

it.	0	1	2	3	4	5	6	7	8	9
SVM	3.0	5.6	**9.2**	**13.2**	**17.6**	**22.2**	**26.9**	**31.6**	**36.4**	**41.2**
NN	3.0	**5.8**	8.8	12.0	15.3	18.7	22.0	25.5	28.9	32.3
browsing	3.0	5.7	8.5	11.2	13.9	16.6	19.2	21.8	24.4	27.0

Table I
PRECISION AND "USER PERCEIVED" RECALL IN THE CASE OF *single concept scenario*.

The other scenarios tested are the *multiple concept AND scenario* and the *multiple concept OR scenario*. We remind to the reader that the *AND scenario* is similar to the "single concept" scenario, where the single concept the user is looking for is actually a combination of concepts. It turns out that this scenario is more difficult than the previous one. For these scenarios we started from each one of the 1.294 queries and tested all the 658 pairs tags as target. Thus, 11.171 retrieval tasks were performed for each relevance feedback method. We made the same comparisons as in the previous scenario, plus we added a comparison on the *concept correlation* measures C_1, and C_2 proposed in Section III for the *AND scenario*. We recall that these measures are useful to understand if, when we are performing a multiple concept retrieval, the process is driven by the combination of concepts rather than the single concepts used in the combination itself.

In Table II the average results in terms of *Precision*, and "user perceived" *Recall* are reported for this scenario. The performance measures exhibit the same behavior as in the case of single tags. With respect to the previous case, the performance are lower because the targets are more difficult to "learn", but it is clear that as the number of iterations increase the relevance feedback shows its advantages also in these difficult tasks. In Figure 1 the average value of the

Precision

it.	0	1	2	3	4	5	6	7	8	9
SVM	12.1	10.7	11.7	**12.9**	**13.8**	**14.7**	**15.4**	**15.9**	**16.3**	**16.7**
NN	12.1	**12.2**	**12.4**	12.5	12.6	12.6	12.6	12.6	12.6	12.6
browsing	12.1	11.3	11.0	10.9	10.7	10.6	10.4	10.6	10.3	10.2

Recall

it.	0	1	2	3	4	5	6	7	8	9
SVM	1.3	2.2	3.6	**5.3**	**7.1**	**9.0**	**10.9**	**12.8**	**14.7**	**16.5**
NN	1.3	**2.6**	**3.9**	5.2	6.5	7.7	9.0	10.2	11.4	12.5
browsing	1.3	2.4	3.5	4.4	5.6	6.6	7.6	8.6	9.6	10.6

Table II
PRECISION AND "USER PERCEIVED" RECALL IN THE CASE OF *multiple concept AND scenario*.

concept correlation measures C_1 and C_2 is reported for all the possible tasks with pairs of tags. All the three relevance feedback techniques exhibit the same behavior. The C_1 measure, computed over all the queries and pairs of tags, shows that only for a small part of pairs it achieves high value, thus meaning that in the majority of cases the feedback have to be submitted according to the combination of the tags/concepts rather than using the single tags/concepts for find images that are relevant with the combination. The C_2 measure tells that in the majority of cases the tags who compose the pair are loosely correlated in the collection used. The results of the *concept correlation* measures allow us to point out that the results shown in Table II are reliable because we have a medium-low correlation in the majority of cases. It means that the combined concept can't be easily retrieved by using only the single concepts, but it is necessary to use the combined concept as feedback in the retrieval process. The values of these two *concept correlation* measures allow us to point out that multiple concepts are actually retrieved by feedback related to multiple concepts rather than by feedback related to individual concepts, and that relevance feedback techniques are effective also in the case of the complex tasks implemented in the *AND scenario*.

In Table III the average results in terms of *Precision*, and "user perceived" *Recall* are reported for the *OR scenario*. With respect to the previous cases, the performance are higher because the targets are more easier: i.e., by recalling what we wrote in Section II at each interaction the images that are relevant to the user's query are those who belong to one of the concepts or to a combination of them, thus at each interaction we have a large number of relevant images.

VI. CONCLUSIONS

In this paper we have proposed "new" scenarios to be used to evaluate a relevance feedback methodologies, based

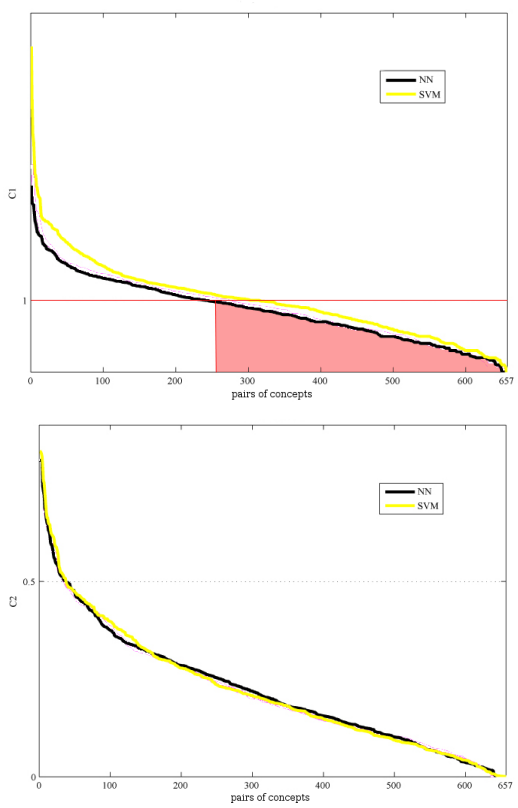

Figure 1. *Concept correlation* measures C_1 and C_2 in the case of *multiple concept AND scenario*.

Precision

it.	0	1	2	3	4	5	6	7	8	9
SVM	50.4	50.3	**54.1**	**57.3**	**59.8**	**61.6**	**63.0**	**64.1**	**65.0**	**65.6**
NN	50.4	**52.1**	53.4	54.3	55.0	55.5	55.9	56.2	56.4	56.6
browsing	50.4	50.0	49.6	49.0	48.7	48.3	47.9	47.4	47.0	46.6

Recall

it.	0	1	2	3	4	5	6	7	8	9
SVM	8.0	12.4	20.1	28.4	37.1	45.9	54.9	63.9	72.8	81.8
NN	8.0	**16.6**	**25.6**	**34.6**	**43.8**	**53.2**	**62.5**	**71.8**	**81.1**	**90.4**
browsing	8.0	14.1	21.2	26.9	33.1	39.8	45.5	50.2	56.8	63.9

Table III
PRECISION AND "USER PERCEIVED" RECALL IN THE CASE OF *multiple concept OR scenario*.

on multi-tagged images. Moreover, two concept correlation measures had been proposed to support the analysis of performance in the case of multiple concept query scenario. The experiments show the performance of two relevance feedback methods in these scenarios, and compared them with simple browsing. The proposed scenarios and measures are to be considered as a starting point to a different way of performing and analyze the behavior of relevance feedback techniques, and future works may expand them.

REFERENCES

[1] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE PAMI*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain, "Content-based multimedia information retrieval: State of the art and challenges," *ACM Trans. Multimedia Comput. Commun. Appl.*, vol. 2, no. 1, pp. 1–19, 2006.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, pp. 1–60, 2008.

[4] T. Huang, C. Dagli, S. Rajaram, E. Chang, M. Mandel, G. Poliner, and D. Ellis, "Active learning for interactive multimedia retrieval," *Proceedings of the IEEE*, vol. 96, no. 4, pp. 648 –667, april 2008.

[5] X. S. Zhou and T. S. Huang, "Relevance feedback in image retrieval: A comprehensive review," *ACM Multimedia Systems*, vol. 8, no. 6, pp. 536–544, 2003.

[6] Y. Rui and T. S. Huang, "Relevance feedback techniques in image retrieval," in *Principles of Visual Information Retrieval*, Springer-Verlag, London, 2001, pp. 219–258.

[7] G. Giacinto, "A nearest-neighbor approach to relevance feedback in content based image retrieval," in *CIVR '07: 6th ACM Int. Conf. on Image and video retrieval*, 2007, pp. 456–463.

[8] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine Learning*, vol. 6, no. 1, pp. 37–66, 1991.

[9] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*. New York: John Wiley and Sons, Inc., 2001.

[10] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander, "LOF: identifying density-based local outliers," *SIGMOD Rec.*, vol. 29, no. 2, pp. 93–104, 2000.

[11] D. Tax, "One-class classification," phd, Delft University of Technology, Delft, June 2001.

[12] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[13] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *Proc. of the 9th ACM Intl Conf. on Multimedia*, 2001, pp. 107–118.

[14] G. Giacinto and F. Roli, "Bayesian relevance feedback for content-based image retrieval," *Pattern Recognition*, vol. 37, pp. 1499–1508, 2004.

[15] L. Piras and G. Giacinto, "Neighborhood-based feature weighting for relevance feedback in content-based retrieval," *Image Analysis for Multimedia Interactive Services, International Workshop on*, vol. 0, pp. 238–241, 2009.

[16] M. J. Huiskes and M. S. Lew, "The mir flickr retrieval evaluation," in *MIR '08: 1st ACM Int. Conf. on Multimedia Information Retrieval*, 2008, pp. 39–43.