Processing complex similarity queries: a systematic approach*

Anna Yarygina[1], Boris Novikov[1], and Natalia Vassilieva[1,2]

[1] Saint-Petersburg University
anya_safonova@mail.ru, borisnov@acm.org
[2] HPLabs
nvassilieva@hp.com

Abstract. In this paper we introduce a set of operations for complex manipulation with approximate queries represented by scoring functions. We show correspondence of these operations to the user expectations and applications needs. We evaluate the effectiveness of our techniques in the context of content-based image retrieval task and compare it to the known fusion methods.

1 Introduction

Both amount and complexity of stored data rapidly grow. The information is presented in the form of complex objects with explicit or implicit structure. Examples may include complex hypermedia objects containing text, images, sound, and video, and dynamic pages typically generated from structured data with attributes of different types. Neither navigational link chasing nor simple keyword queries cannot be sufficient for advanced search or complex object retrieval.

The reasons to evaluate complex structured queries are:

- A need to combine search criteria for different types of information;
- A query refinement, e.g. based on user profile or feedback;
- Advanced users may need query structuring.

Database query languages, e.g. relational, are nearly perfect when an exact result is expected. However, in the world of similarity-based queries with approximate results the behavior of logical operators does not always meet human expectations. For example, logical operators cannot capture the expected increase in confidence due to multiple sources of support.

Approximate similarity-based search results are usually represented by scores of objects. A concept of data fusion is suitable to combine results of several similarity-based queries. Several data fusion methods were proposed and empirically studied in different information retrieval environments and showed good results. These methods require appropriate calibration of incoming scores.

* This research is supported by RFBR, grant 10-07-00156.

In this paper we introduce a systematic approach to construction of complex similarity queries. We define a model for complex queries and proceed with specification of several calibration and fusion operations. We evaluate the proposed techniques using content-based image retrieval environment and show how to use these operations for semi-automatic calibration of incoming scores. The results are comparable to the best ones obtained with empirical approaches.

The paper is structured as follows. The model overview in section 2 is followed by its specification in section 3. The experiments and the analysis of their results are presented in section 4. Section 5 outlines the related work.

2 Similarity Query Model

2.1 Similarities and Distances

Modern information retrieval models assume that relevance of an object to a query is expressed as similarity. Typically an object and a query are represented as feature vectors, while similarity is calculated as a similarity measure (e.g. cosine measure) or via distance function (e.g. Euclidian, l_1). Although both feature extraction techniques and similarity functions are crucial for quality and effectiveness of the retrieval systems, these issues are out of the scope of our research.

In our work we define the similarity measure and distance function as follows.

Definition 1 *Similarity measure on the set S is the function* $\mathsf{sim} : S \times S \rightarrow [0, 1)$*:* $\forall a, b \in S\ \mathsf{sim}(a, a) = 1$ *and* $\mathsf{sim}(a, b) = \mathsf{sim}(b, a)$.

Several similarity measures may be needed simultaneously on the same set.

Definition 2 *Distance function on the set S is the function* $\mathsf{dist} : S \times S \rightarrow R$: $\forall a, b \in S\ \mathsf{dist}(a, b) \geq 0, \mathsf{dist}(a, b) = \mathsf{dist}(b, a), \mathsf{dist}(a, a) = 0$; *that is a semi-metric.*

The notions of distance and similarity are closely related and similarity measure can be constructed from distance (similaring) and vice versa (distancing). We use the following formula for this conversion: $\mathsf{sim}(a, b) = 1/(1 + \mathsf{dist}(a, b))$.

2.2 Queries and Result Sets

Any querying system might return either exact or approximate result. The former is usually is a set of objects satisfying the query, while the latter is typically a ranked list based on objects' scores. We describe a model which captures both query paradigms and allows combining different types of subqueries in a single request. The central concept of this model is a Q-set. It abstracts from query language, nature of objects, or query evaluation technique, and encapsulates both a query and the result of its evaluation. The notion of the Q-set essentially coincides with the notion of the fuzzy set.

Definition 3 *Q-set=(S,score), where S is a set of objects, $score : S \to [0,1]$ is a scoring function. $\mathcal{Q}(S)$ is the set of Q-sets defined on S.*

Fuzzy set-theoretic operations can operate on Q-sets. However, these operations do not possess certain desirable properties important for information retrieval. Thus, we have to introduce additional operations in our model.

The concept of the Q-set can accommodate various types of queries, e.g. exact database queries and queries in probabilistic data bases. For every object in S we can construct a Q-set from a similarity function defined on S. Function $\mathsf{scoring}$ by object $q \in S$ and a similarity measure sim defines $A \in \mathcal{Q}(S)$ such that $A.score(x) = \mathsf{sim}(q, x)$.

As soon as several Q-sets are constructed on the same set S using either similarity measures based on objects' features or exact queries, we can manipulate with Q-sets regardless of how they were produced.

A generic operation which combines several Q-sets into one is known as fusion. In this research we explore properties of different fusion techniques.

Intuitively expected fusion properties identified in [7] are the *Chorus* and *Skimming* Effects. The former suggests that objects with high scores in both arguments should be preferred in the result, while the latter promotes objects with high scores in at least one argument. Depending on the application needs, the fusion operation may take these effects into account differently. Thus, no single implementation can be the best for all applications.

In our model an operation has the Chorus effect if the output score of an object is higher than its scores in both argument Q-sets.

2.3 Operations

All operations introduced in our model are inside the class of Q-sets. Thus the class of Q-sets is closed under the proposed set of operations.

The definition 4 lists the basic requirements for a generic fusion operation. In addition, it may possess other desirable properties such as Chorus and Skimming Effects.

Definition 4 *The fusion operation is a function* $\mathsf{fusion} : \mathcal{Q}(S) \times \mathcal{Q}(S) \to \mathcal{Q}(S)$*:* $\mathsf{fusion}(A, B) = \mathsf{fusion}(B, A)$*; preserves order ($A.score(x) < A.score(y)$, $B.score(x) < B.score(y) \Rightarrow \mathsf{fusion}(A, B).score(x) < \mathsf{fusion}(A, B).score(y)$).*

The fusion can be effective only if its arguments are comparable, obviously not the case for arbitrary Q-sets. Scoring functions constructed from different features and similarity measures may vary in range and distribution. Thus, the incoming arguments must be calibrated to ensure their scores are comparable.

Informally, scoring functions are comparable if the scores of important objects do not differ too much. Typically the objects with high scores are important.

We introduce two operations to be used in the calibration procedures: the normalization and strengthening. The former re-scales the scores or distances evenly, while the latter increases high scores and decreases low ones. We use norm to denote the normalization operation.

Definition 5 *The strengthen operation is a monotonic function* $\mathsf{strengthen} : \mathcal{Q}(S) \times [0,1] \rightarrow \mathcal{Q}(S)$*:* $\forall l \in [0,1]$, $\forall x \in S$, $\forall A \in \mathcal{Q}(S)$

$$\mathsf{strengthen}(A,l).score(x) \begin{cases} \geq A.score(x) \ , |y : A.score(y) \geq A.score(x)| \leq l * |S| \\ \leq A.score(x) \ , |y : A.score(y) \geq A.score(x)| \geq l * |S| \end{cases}$$

The inverse operation is called weaken*.*

The factors to be taken into account in the design of calibration algorithms and underlying operations norm, strengthen, and weaken are:

Stability Outliers should not affect the calibration significantly.
Skew Higher scores should provide more impact.
Effectiveness Argument Q-sets of different quality should differ in the impact.

The last item suggests that a priori knowledge, such as relative precision of different features, should be used to weaken or strengthen Q-sets appropriately. For example, if texture features of an image result in high scores but less precision than color ones, then the former should be weakened. Such a priori knowledge may be obtained either from expert or from training data sets with machine learning procedures.

3 Refinement

In this section several alternative implementations are specified for each operation defined above. Algorithms and formulas which conform with intuitive expectations are presented. Each implementation meets the requirements presented in the section 2 and provides some specific properties.

3.1 Unary Operations

Normalization Several alternatives of the norm operation are specified in table 1. Each operation takes a Q-set $A \in \mathcal{Q}(S)$ as an argument.

Table 1. Implementations of calibration operations

Operation	Formula
norm-maxmin	$\frac{A.score(e)-min(A.score(x))}{max(A.score(x))-min(A.score(x))}$
norm-avg	$\mathsf{similaring}(\frac{\mathsf{distancing}(A.score(e))}{avg(\mathsf{distancing}(A.score(x)))})$
norm-dist(α)	$\mathsf{similaring}(\mathsf{distancing}(A.score(e)) * \alpha)$

The first normalization algorithm (norm-maxmin) is usually applied in the context of fusion. This algorithm is used to change the range of scoring functions'

values to [0,1]. According to definition 3 scoring functions already limited to such range. Nevertheless the advantage of this normalization algorithm is that values of scoring function always fill the whole range [0,1].

By definition the norm-avg operation brings the average value of distance function to 1. This implementation of normalization operation is not sensitive to the outliers contrary to norm-maxmin. The operation norm-dist provides simple calibration technique which can be applied in more complex scenarios.

Calibration In order to support the skew property defined in section 2, we developed a procedure enabling fine-tuned calibration of scoring functions to make the distributions of scores in Q-sets comparable. Scoring functions are normalized by making equal distances at certain level. The level represents the threshold which splits the most important scores from others.

The parameter $p \in [0,1]$ defines a portion of objects with highest scores, which are considered important. Procedure $\mathsf{normalize\text{-}dist}_p$ constructs $A \in \mathcal{Q}(S)$ from $A_1, A_2 \in \mathcal{Q}(S)$:
$\forall e \in S\ A.score(e) = \mathsf{norm\text{-}dist}(\alpha)(A_1).score(e)$, where $\mathsf{norm\text{-}dist}(\alpha)$ is defined in table 1, $\alpha = \mathsf{distancing}(A_2.score(x))/\mathsf{distancing}(A_1.score(y))$, x, y: $|z : A_2.score(z) \geq A_2.score(x)| = p * |S|$ and $|z : A_1.score(z) \geq A_1.score(y)| = p * |S|$.

Weaken and Strengthen Strengthen and weaken operations are needed during the Q-sets fusion to exploit a priori knowledge regarding the quality of the scoring functions.

Definition 6 *Operation* $\mathsf{strengthen(n)}$ *by* $A \in \mathcal{Q}(S)$ *and parameter level* $\in [0,1]$ *constructs* $A_0 \in \mathcal{Q}(S)$*:*
$\forall e \in S\ A_0.score(e) = \mathsf{similaring}((\mathsf{distancing}(A.score(e))/M)^n)$*, where* $n > 1$ *is parameter of procedure, and* $M : |y : \mathsf{distancing}(A.score(y)) \leq M| = level * |S|$.

Operation $\mathsf{weaken(n)}$ *is defined as* $\mathsf{strengthen(1/n)}$.

By definition 6 high scores will decrease and low scores will increase in a Q-set after application of weaken operation. Behavior of weaken and strengthen operations essentially depends on distribution of distance function values with respect to threshold *level*. For example, after normalization according to definition of norm-avg operation, strengthen operation with *level* such that $M = 1$ increases scores which values are higher than average value inside this Q-set.

3.2 Binary Operations

Several alternatives for the fusion operation (see definition 4) are specified in table 2. Each operation takes two Q-sets $A_1, A_2 \in \mathcal{Q}(S)$ as arguments and constructs a new Q-set with scoring function defined in column Formula.

The advantage of super-union and super-intersect is the ability to capture the probabilistic properties at least if the arguments are independent. However

Table 2. Implementations of fusion operations

Operation	Formula	Chorus Effect	Skimming Effect
intersect	$min(A_1.score(e), A_2.score(e))$	-	-
union	$max(A_1.score(e), A_2.score(e))$	-	+
super-intersect	$A_1.score(e) * A_2.score(e)$	-	-
super-union	$1 - (1 - A_1.score(e)) * (1 - A_2.score(e))$	+	+
CombMNZ	$(A_1.score(e) + A_2.score(e)) * R$	+	+

scores produced by super-intersect might be unreasonably low in the context of fusion.

CombMNZ is reported as the best fusion algorithm in various IR environments and possesses nice properties, such as the Chorus Effect, especially in fusion of several queries [3]. We use CombMNZ as a baseline in out experiments.

4 Experiments

To justify our model we conducted a series of experiments in the context of content-based image retrieval task.

4.1 Experimental Environment and Setup

Experimental image database consists of 1087 images. It includes 101 images from Corel Photo Set collection which are used as queries. The query images are divided into 16 groups of similar images by 2 experts. During the experiments images from the same group as the query-image are treated as relevant, while others are not. Every image in the database is represented by a feature vector in three different feature spaces: 1) *color moments* – moment based color distribution features and color metrics from [9]; 2) *color histograms* – color histogram with spatial information encoded into color index with corresponded distance function [10]; 3) *texture* – convolutions of image with ICA filters as a texture feature and Kullback-Leibler divergence as a texture metrics [2].

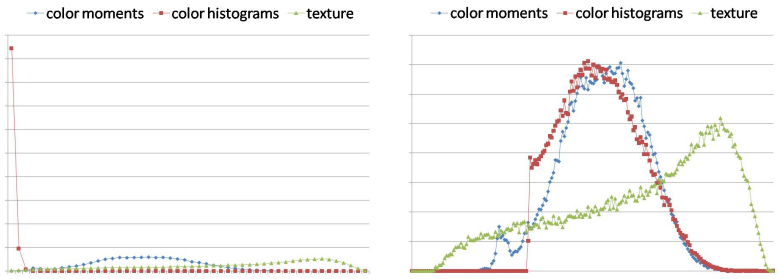
The fusion technique CombMNZ is chosen as a baseline. The applicability and accuracy of proposed fusion operations will be analyzed based on comparison of results obtained by our model and this baseline algorithm. In all our experiments we measure R-precision fusion results relative to R-precision of the input Q-sets.

4.2 Analysis of Experiments

The table 3 shows the range of distance values in our data set. These measurements approve that distance functions corresponding to various types of Q-sets differ from each other. Hence, a calibration is ultimately needed.

Table 3. Distance and precision for color and texture Q-sets

Type of Q-set	min distance	max distance	**R-precision**
colour moments	0.02	6.79	**0.48**
colour histograms	0.22	199.50	**0.40**
texture	0.003	13.47	**0.15**



(a) Normalization by norm-maxmin (b) Normalization by norm-avg

Fig. 1. The distribution of scores

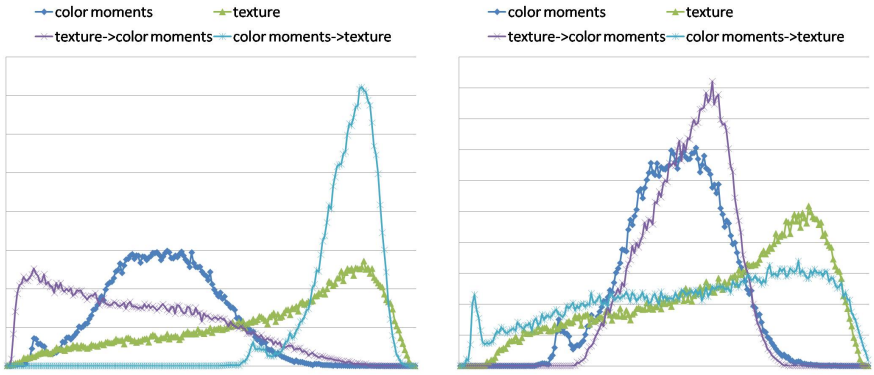
The figure 1 shows the score distributions in different Q-sets after normalization by norm-maxmin (1(a)) and norm-avg (1(b)). Normalized by norm-maxmin and unnormalized Q-sets do not differ significantly.

The figure 2(a) presents the distribution of scoring function values normalized by procedure normalize-dist$_p$ with $p = 0.1$. We assume that in our dataset top 10% of objects significantly influence the results selected after fusion.

We investigate the impact of strengthen and weaken operations in experiments where we apply operation norm-avg to the Q-sets and then operation strengthen. The value of parameter n in strengthen(n) is taken to make equal scores from pair of Q-sets at the level 10%. We further refer to this procedure as norm&strenthen. Histogram 2(b) shows the distributions of scores based on color moments and texture after such processing.

We have measured the R-precision of results produced by separate Q-sets without fusion to analyze the quality improvements obtained by our technique. Table 3 outlines R-precision obtained by using Q-sets based on color moments, color histograms and texture separately. Results presented in table 3 show that texture based features expose poor precision. This a priori knowledge suggests that texture should be weakened.

Table 4 demonstrates R-precision obtained by fusion of two Q-sets. It shows that in most cases in spite of normalization operation super-union gives poor R-precision in the fusion of Q-sets based on texture and color moments. The reason is that one of the Q-sets is defined by non-effective scoring function itself and which dominates even after normalization. The only case when super-union appropriately takes into account scores is norm&strenthen. The histogram pre-



(a) Normalization by norm-prop$_{0.1}$ (b) Normalization by norm-avg&strengthen

Fig. 2. The distribution of values of color moments and texture scoring functions

sented on the figure 2(b) shows the dominance of high scores constructed by color moments. Similar observations are the reason for poor results obtained after application of normalize-dist(0.1) for all fusion techniques. The quality of normalization highly depends on parameters and optional strengthen of those Q-sets which are known to be good for specific retrieval task.

The results obtained by fusion of Q-sets based on texture and color histograms show the sensitivity of operations CombMNZ and super-union to the implementation of norm operation.

The third part of table 4 demonstrates R-precision obtained by fusion of two Q-sets: based on color histograms and color moments. One may see that appropriate normalization technique can improve the retrieval performance. However the obtained results dramatically depend on the quality of the initial Q-sets. In case of the fusion of Q-sets constructed by color moments and color histograms different implementations of fusion operation provide comparable R-precision.

5 Related Work

Our model for complex query processing is closely related to advanced ranking techniques, query languages and data fusion.

Authors of [12] discuss the probability ranking principle proposed by Robertson in 1977, analyze its weakness in the task of multimedia retrieval. They present ranking approach which provides the possibility to take into account both the relevance probability and document transmission and inspection time. Ranking for structured documents is introduced in [11]. An approach to results ranking and weighting for the interactive retrieval is presented in [6]. Most of the complex ranking algorithms rely on a data type.

Table 4. R-precision for fusion results

	without norm	norm-avg	norm-mimmax	norm-dist(0.1)	norm&strenthen
Texture and color moments					
ComMNZ	0.45	0.44	0.46	0.38	0.45
Super-intersect	0.46	0.45	0.47	0.38	0.45
Super-union	0.32	0.33	0.31	0.37	0.46
Texture and color histograms					
ComMNZ	0.20	0.41	0.20	0.37	0.43
Super-intersect	0.45	0.41	0.45	0.37	0.43
Super-union	0.16	0.35	0.16	0.37	0.44
Color moments and histograms					
ComMNZ	0.50	0.53	0.49	0.53	0.53
Super-intersect	0.50	0.53	0.50	0.53	0.53
Super-union	0.49	0.52	0.50	0.53	0.52

The expressiveness and simplicity of the query language influence the quality of search. An extension of relational algebra close to our model and supporting similarity queries is presented in [4].

A discussion of fundamentals of retrieval results' fusion, e.g. the "Chorus Effect", the "Skimming Effect", and the "Dark Horse Effect" are outlined and described in [7]. In [1] a meta search model based on an optimal democratic voting procedure is described and investigated. Authors of [3] consider several algorithms for fusion of multiple document lists. CombMNZ have shown better results on selected experimental data set.

The authors of [5] analyze effective fusion technique for video retrieval. Several fusion strategies based on ranks and scores are compared. The experiments show that appropriateness of fusion technique highly depends on the specific task. In general, the choice of fusion techniques depends on specific collection of data [7, 5, 3]. Fusion algorithms which enable accounting hierarchical structure of retrieved documents are presented in [8].

A probabilistic approach to data fusion called probeFuse is presented in [7]. Authors show that the performance of fusion algorithm can be significantly improved with calibration based on ranges and the reliability of data sources.

The normalization of scores is critical for the quality of any fusion technique and was studied intensively. A normalization by deflection of a score from the minimum one is discussed in [3, 5]. Several alternative normalization techniques are discussed in [13].

6 Conclusions

In this paper we introduced a systematic approach to construction of approximate complex queries represented as scoring functions. We define query calibra-

tion and fusion algorithms which meet the high-level semantic expectations and provide for consistent probabilistic interpretation of resulting scores.

The experiments clearly show that proposed techniques can be configured to provide predictable results. The effectiveness of our techniques is comparable to known algorithms. The proposed techniques are useful for querying complex objects against their structure, textual and multimedia content. Our operations provide a base for modeling complex querying scenarios such as relevance feedback, shuttle search, and combination of structured and unstructured retrieval.

References

1. J. A. Aslam and M. Montague. Models for metasearch. In *Proc. 24th SIGIR*, SIGIR '01, pages 276–284, New York, NY, USA, 2001. ACM.
2. H. Borgne, A. Guerin-Dugue, and A. Antoniadis. Representation of images for classification with independent features. *Pattern Recognition Letters*, 25:141–154, 2004.
3. A. Chowdhury, O. Frieder, D. A. Grossman, and M. C. McCabe. Analyses of multiple-evidence combinations for retrieval strategies. In W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, editors, *SIGIR*, pages 394–395. ACM, 2001.
4. P. Ciaccia, D. Montesi, W. Penzo, and A. Trombetta. Imprecision and user preferences in multimedia queries: A generic algebraic approach. In *Proc. FoIKS '00*, pages 50–71, London, UK, 2000. Springer-Verlag.
5. K. Donald and A. Smeaton. A comparison of score, rank and probability-based fusion methods for video shot retrieval. In *Image and Video Retrieval*, volume 3568 of *LNCS*, pages 61–70. Springer Berlin / Heidelberg, 2005. 10.1007/11526346-10.
6. N. Fuhr. A probability ranking principle for interactive information retrieval. *Inf. Retr.*, 11(3):251–265, 2008.
7. D. Lillis, F. Toolan, R. W. Collier, and J. Dunnion. Probfuse: a probabilistic approach to data fusion. In *SIGIR*, pages 139–146. ACM, 2006.
8. S. Shi, B. Lu, Y. Ma, and J.-R. Wen. Nonlinear static-rank computation. In D. W.-L. Cheung, I.-Y. Song, W. W. Chu, X. Hu, and J. J. Lin, editors, *CIKM*, pages 807–816. ACM, 2009.
9. M. Stricker and A. Dimai. Spectral covariance and fuzzy regions for image indexing. *Mach. Vision Appl.*, 10(2):66–73, 1997.
10. N. Vassilieva. *Content-based image retrieval methods.* PhD thesis, Saint Petersburg State University, 2010.
11. J.-N. Vittaut and P. Gallinari. Machine learning ranking for structured information retrieval. In *ECIR*, volume 3936 of *LNCS*, pages 338–349. Springer, 2006.
12. M. Wechsler and P. Schäuble. A new ranking principle for multimedia information retrieval. In *Proceedings of the fourth ACM conference on Digital libraries*, DL '99, pages 146–151, New York, NY, USA, 1999. ACM.
13. S. Wu and S. McClean. Performance prediction of data fusion for information retrieval. *Inf. Process. Manage.*, 42:899–915, July 2006.