A Flexible Information Retrieval Method for Bird Databases

Yo-Ping Huang[1], Tienwei Tsai[2], Mann-Jung Hsiao[3]

Department of Computer Science and Engineering
Tatung University
Taipei, Taiwan 10451 R.O.C.
yphuang@ttu.edu.tw[1], twt@mail.chihlee.edu.tw[2], hsiao@mis.knjc.edu.tw[3]

Abstract

The information retrieval techniques should provide users with easy access to the information in which they are interested. Unfortunately, characterization of the user information need is not a simple problem. In ecological databases, for example, it is very hard to formulate a query for a new learner or an inexperienced user. To overcome such problems, we have proposed an information retrieval method that allows users to conduct a query by transforming human perceptions into numerical data based on their impressions or opinions on the target. However, the former system gave a fixed weighting value for such an individual feature. Based on this basic idea, we propose some further improvements in this paper. A set of weighting vectors is provided to enable the user to express a measure of the relative importance among the features if he/she can not be reasonably sure about some of them. We develop a bird searching system to demonstrate the effectiveness of the proposed method. The experimental results show that the system introduces great flexibility with the assistance of weighting vectors.

Keywords: Fuzzy semantic query, Information retrieval, Weighting vectors.

1. Introduction

Keyword-based queries are popular because they are intuitive, easy to express, and allow for fast ranking. However, the simplicity of the approach prevents the formulation of more elaborated querying tasks [1]. For many kinds of data, in practice, it is difficult to conduct a query with a complete set of keywords. Moreover, even all the data are characterized and annotated, difficulties may still arise because users are likely to express the impressions from different angles and at different levels of perceptions. Especially, this problem is often confronted while retrieving information from an ecological database. Due to the inexact and uncertain perceptions of users, data modeling should be intentionally designed to support fuzzy data content, structure, and presentation. It must also allow users to conduct a query by the fuzzy perceptions rather than by exact keywords. For those queries without explicitly giving the keywords or with the semantic contents that cannot be translated into crisp or clear forms, we call such queries the fuzzy semantic queries.

Some research works have been dedicated to the bird information retrieval [2-5]. In [2], the proposed information extraction technique focused on building a bird database from text information. The concepts of entropy and data-missing rate are used to automatically choose the features for characterizing a bird. Another work by Chen et al. developed a mobile scaffolding-aid-based bird-watching learning system, which aimed to construct an outdoor mobility-learning activity under the up-to-date wireless technology [3]. The work in [4] incorporated an alternative teaching methodology into a curriculum. In this course, students were asked to build a bird searching system on PDAs (Personal Digital Assistants), but the query was simply based on keyword approach.

The systems mentioned above did not support fuzzy semantic queries. The fuzzy semantic approach for retrieving bird information was first proposed in [5]. The formulation of a query was done by transforming human perceptions into numerical data. It did a good job in capturing users' perceptions. However, the weighting vector which represents the importance of individual feature is created at the beginning of the system on probability basis. It did not allow users to express their opinions or impressions on an individual feature. In this paper, we provide pictorial feature examples and a friendly user interface for users to specify the weighting vector based on their observations.

This paper is organized as follows. The former system that supports fuzzy semantic query is introduced in Section 2. The proposed flexible weighting vectors are illustrated in Section 3. In section 4, we examine the effectiveness of our proposed method. Section 5 concludes this paper.

2. Fuzzy Semantic Query Model

The classic models in information retrieval consider that each document or object is described by a set of representative keywords called index terms. In the fuzzy semantic query model, however, each instance is represented by integrating two content descriptors: centrality and intensity, each of which represents an abstract level of similarity. Before getting into more details of them, we define some basic components.

Definition 1. The fuzzy semantic query can be modeled as [**D**,**Q**,**F**(q, d_x)], where

(1) **D** represents a set of instances in the database.

(2) **Q** is a set of query vectors for the user information needs.

(3) **F**(q, d_x) is a ranking function which defines the dissimilarity between the query q and the instance d_x, where $q \in$ **Q** and $d_x \in$ **D**.

In this model, we do not intend to eliminate the semantic ambiguity neither in database nor in user query. On the contrary, we explicitly represent and process ambiguity by introducing two content descriptors, which will be illustrated in the following.

Definition 2. The instance vector and query vector are both composed of a centrality vector and an intensity vector, where

(1) **centrality vector** ***C*** is a set of terms that provide the quantification of class or category similarity. Let ***C*** = (C_1, C_2, …, C_m), where m is the number of category and C_i denotes the similarity between the instance and category i.

(2) **intensity vector** ***I*** is a set of terms that provide the quantification of characteristic or feature similarity. Let ***I*** = (I_1, I_2, …, I_n), where n is the number of features that characterize an object. Each feature I_j will have p_j terms if there are p_j options (or attributes) for the feature j. Each term in I_j indicates its similarity degree for that corresponding option.

Therefore, an instance in the database or a user query can be represented by integrating its category centrality vector ***C*** and its feature intensity vector ***I***.

Definition 3. An instance vector d_x and a query vector q are formed as $d_x = (C^{(d_x)}, I^{(d_x)})$ and $q = (C^{(q)}, I^{(q)})$, respectively. Note that $C^{(d_x)}$ and $I^{(d_x)}$ are the centrality vector and the intensity vector of instance d_x, respectively. Similarly, $C^{(q)}$ and $I^{(q)}$ are the centrality vector and the intensity vector of a query q, respectively.

Figure 1 shows the five categories used in our experimental system and Table 1 lists all the bird features and their attribute values in the system.

Fig. 1: The main screen of the system.

Table 1. Bird features used in the system.

Feature	Attribute values (options)
Body size (I_1)	Bigger than a sparrow, Similar or smaller than a sparrow.
Beak shape (I_2)	Duck type, Long type, Hooked type, Short type.
Tail shape (I_3)	Long and forked, Long and unforked, Short and forked, Short and unforked.
Flying way (I_4)	Wave, Straight, Spiral.
Walking way (I_5)	Jump, Walk.
Feather color (I_6)	Dark, Light, Red, Blue, Green.
Foot length (I_7)	Long, Short.

Example 1. A centrality vector ***C*** = {0.2, 0, 0, 0.8, 0} means that the instance has 20% of similarity with the first category, 80% of similarity with the fourth category, and no similarity with the remaining categories.

Example 2. A vector I_1 = {0.8, 0.2} denotes the user has 80% of confidence that the observed bird is bigger than a sparrow. Similarly, a vector I_2 = {0, 0.7, 0.3, 0} means the user has 70% of confidence for the long beak and 30% of confidence for the hooked beak for a given instance.

3. The Weighting Vectors

When users intend to retrieve desired information from bird database, a more likely scenario is one in which the similarity and difference of features to be described are not obvious. To overcome this problem, we involve two weighting factors w_1 and w_2 to indicate the significant levels of centrality and intensity, respectively. The values are application dependent and can be specified according to the opinions of users. To allow users express the confidence level for each feature while formulating a query, we involve another weighting vector **R**. **R** is in the form of (r_1, r_2, …, r_n), where r_j is the weight for feature *j*.

Example 3. Intuitively, a weighting vector R can be set to (0.5, 0.25, 0.25, 0.33, 0.5, 0.2, 0.5). Each term is the inverse of the number of options in that feature. The feature "beak shape" has four options: duck type, long type, hooked type, and short type. Therefore, r_2 equals to 0.25.

The vector obtained above is called probability-based vector. Each term in the vector equals to the probability of a correct guess if a user has no idea about the associated feature. We tend to allow users to assign a weight for each feature. Thus, the feature with the highest value is considered the most discriminating feature in describing a query. For example, if the query object is special in its color feature, the color can be used as the main features while other features as auxiliary features.

Basically, we adopt Euclidean distance to calculate the dissimilarity between a query and instances in the database. The Euclidean distance function between two vector V_1 and vector V_2 is known as:

$$S(V_1,V_2)=(\sum_{k=1}^{t}(V_1[k]-V_2[k])^2)^{1/2}, \qquad (1)$$

where both V_1 and V_2 have t terms. Notation V_1[k] stands for the kth term of V_1 and V_2[k] stands for the kth term of V_2. The measuring cost is thus highly subject to the number of instances in the database. Since the size of the database is almost fixed, both the computational complexity and the memory requirements are under control.

Definition 4. The dissimilarity measure function F of an instance d_x and a query vector q is defined as

$$F(d_x,q)=w_1*S(C^{(d_x)},C^{(q)})+ w_2*(\sum_{j=1}^{n}r_j*S(I_j^{(d_x)},I_j^{(q)})), \qquad (2)$$

where r_j is the weight for the *j*th feature. In calculating the distance between d_x and q, $S(C^{(d_x)},C^{(q)})$ denotes the centrality distance and $S(I_j^{(d_x)},I_j^{(q)})$ represents the intensity distance for the *j*th feature.

Note that a normalization for the weighting vector **R** should be taken before calculating the distance. That is

$$r_j=r_j/(\sum_{j=1}^{n}r_j). \qquad (3)$$

4. Experimental Results

To gain insight into the benefits of weighting vectors, we develop a bird searching system to investigate their effects. Fig. 1 is the main screen of our system. Fig. 2 is a query bird observed by a user. Users can use a friendly user interface to input all the observed characteristics, and then the system will formulate and transform them into a query vector represented by centrality degrees and intensity degrees as depicted in Definition 2 and Definition 3. Besides, the user can assign two weighting factors w_1 and w_2 to indicate the significant levels of centrality and intensity. And again, the user can further assign another weighting vector **R** to emphasize or decrease each individual feature. The system will initially give a weighting vector **R** based on the probability of each attribute value as given in Example 3. To make the system more flexible, it allows users to express their opinions by modifying the weighting vector. If a feature is considered to be significant, its associated weight should reflect the true importance. The removal or degrading of uncertain features from a query is another way to improve the sensitivity of similarity measurement performed with that query. The resulting distance is calculated by adding the centrality distance and intensity distance between the query example and contents in the database as shown in equation (2).

Fig. 2: A query example.

The fuzzy semantic query model usually doesn't identify a perfect match because the query is inherently imprecise. For this reason, the system will output the top five probable matches listed by the descendent order of their match degrees from left to right. It is also usually impossible to prove that the results obtained through these weighting vectors are the best obtainable because the user plays an important role in judging the query quality. In our case, though the given query example satisfies a user, it might have

different results for another example. The most important issue here is that the system does allow users to distinguish the importance among features based on their opinions. Table 2 is an input example for Fig. 2 and Table 3 shows the default weighting vector and the modified one according to a user's opinions. Fig. 3a is the search result on the probability basis while Fig. 3b is the search result on the user's opinions. It is observed that the target bird is moved forward from rank 4 to rank 2.

Table 2. An input example for Fig. 2.

Category/Feature	Attribute values (options)
Category	(0, 0, 90, 10, 0)
Body size (I_1)	(0, 100)
Beak shape (I_2)	(0, 0, 50, 50)
Tail shape (I_3)	(0, 0, 50, 50)
Flying way (I_4)	(33, 33, 33)
Walking way (I_5)	(100, 0)
Feather color (I_6)	(90, 10, 0, 0, 0)
Foot length (I_7)	(0, 100)

Table 3. The weighting vector **R**.

	Weighting vector **R**
Default (probability)	(0.5, 0.25, 0.25, 0.33, 0.5, 0.2, 0.5)
User's opinions	(1.0, 0.25, 0.25, 0.33, 0.5, 1.0, 1.0)

Fig. 3a: The result based on the default weighting vector.

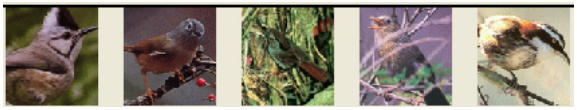

Fig. 3b: The result based on the user's weighting vector.

5. Conclusions and Future Works

In this paper, we first give a brief description of a fuzzy semantic query model and then provide a flexible user interface for a user to generate his/her query. This model has already been proved to be effective for searching bird databases in handheld devices that cannot support easy input of query. In our experimental system, we allow users to specify the query object with two major content descriptors: centrality and intensity. Since several descriptors are used simultaneously, it is necessary to integrate similarity scores resulting from the matching processes in different feature spaces. Two sets of weighting factors are used to improve the drawbacks due to the uncertainties of users. The first set, i.e., w_1 and w_2, is used to indicate the relative significance levels of the centrality and the intensity. The second one, i.e., R, is used to emphasize the implicit degrees of importance among features. It is potentially useful when the user cannot be reasonably sure about this feature. With the assistance of our user interface, he/she can easily decrease the weighting to its effect. On the other side, the user can also increase the weighting for the feature in which he/she is the most confident. Our system shows a very promising result because the input procedure of weighting vectors is rather simple and appears to be quite reasonable.

Though this experimental system is proved to be generic and flexible by involving some weighting vectors, there is a room for more effectiveness. The feature weighting vector used in the system is fairly subjective. We intend to derive a more elaborated weighting vector from the existing database in the future. Besides, we want to investigate the effects of an instance belonging to multiple classes.

Acknowledgments

This work is supported by National Science Council, Taiwan, R.O.C. under Grants NSC92-2213-E-036-017, NSC92-2516-S-036-001, and by Tatung University under Grant B9208-I02-025.

References

[1] R. Baeza-Yates and B. Ribeiro-Neto, *Modern Information Retrieval*, Addison Wesley, Boston, MA, 1999.

[2] C.-H. Chang, D.-S. Wu, P.-K. Shih, Y.-H. Lin, and Y.C. Chen, "Information extraction and query model design for Taiwan wild bird database," in *Proc. The Joint conference on AI, Fuzzy System and Grey System*, Taipei, Taiwan, Dec. 2003.

[3] Y.-S. Chen, T.-C. Kao, J.-P. Sheu, and C.-Y. Chiang, "A mobile scaffolding-aid-based bird-watching learning system," in *Proc. of the IEEE Int. Workshop on Wireless and Mobile Technologies in Education*, pp.15-22, 2002.

[4] Y.-P. Huang and T.-W. Chang, "Using handheld devices as alternatives to inspire students innovation in designing new information technology," in *Proc. Int. Conf. on Engineering Education*, Valencia, Spain, pp.4057-1-4057-8, July 2003.

[5] Y.-P. Huang, L.-J. Kao, T. Tsai, and D. Liu, "Using fuzzy centrality and intensity concepts to construct an information retrieval model," in *Proc. IEEE Int. Conf. on Systems, Man & Cybernetics*, Washington, D.C., pp.3257-3262, Oct. 2003.