Similarity Between Multi-valued Thesaurus Attributes: Theory and Application in Multimedia Systems

Tom Matthé[1], Rita De Caluwe[1], Guy De Tré[1], Axel Hallez[1], Jörg Verstraete[1], Marc Leman[2], Olmo Cornelis[2], Dirk Moelants[2], and Jos Gansemans[3]

[1] Ghent University, Dept. of Telecommunications and Information Processing, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium
{Tom.MATTHE, Rita.DECALUWE, Guy.DETRE, Axel.HALLEZ, Jorg.VERSTRAETE}@UGent.be
[2] Ghent University, Dept. of Musicology, Blandijnberg 2, B-9000 Gent, Belgium
{Marc.LEMAN, Olmo.CORNELIS, Dirk.MOELANTS}@UGent.be
[3] Royal Museum for Central Africa, Dept. of Ethnomusicology, Leuvensesteenweg 13, B-3080 Tervuren, Belgium
Jos.GANSEMANS@africamuseum.be

Abstract. In this paper, the theoretical aspects of calculating the similarity between sets, and its generalizations multisets, fuzzy sets and fuzzy multisets, is presented. Afterwards, this theory is applied to enhance the facilities for accessing a multimedia system, namely when searching for correspondence between multi-valued attributes, which are coupled with a thesaurus. Furthermore, to allow flexibility in this search, thesauri with similarities defined between the thesaurus terms are considered. As a possible application, the DEKKMMA project is introduced, a project about an audio archive of African music.

1 Introduction

Musical audio is an important multimedia type. It is also an important issue for archivists all over the world. The DEKKMMA-project[1], a cooperation between the Royal Museum for Central Africa and Ghent University, is an example of this. Mainly due to Belgium's colonial history, the Royal Museum for Central Africa near Brussels owns one of the largest and world wide most important collections of music recordings from Central Africa. Apart from conservation and sometimes restoration of the historical sources such as wax cylinders, sonofil wires and magnetic tapes, the museum wanted to preserve its rich musical audio collection and linked contextual information in a digital audio archive. The primary goal of the DEKKMMA-project is to digitize the museum collection and to set up a database

[1] The DEKKMMA-project is financed by the Belgian Federal Science Policy Office under project nr. 12/AE/212. The work presented in this paper is realized within the scope of this project.

H. Legind Larsen et al. (Eds.): FQAS 2006, LNAI 4027, pp. 331–342, 2006.

system that allows to store both the digital sources and the contextual information. The secondary goal of the project is to open up the archive to a broad public in such a way that different groups of users, such as visitors, researchers, etc., can easily and efficiently find the information they are looking for.

This paper focusses on one aspect of accessing such a multimedia system, namely finding similarities between attributes of the meta-data[2] of different records in the database, such as the title of the piece, the geographical origin, the name of the musicians, instruments used ... This is an important aspect when searching for similarities between different records, e.g. when querying by 'example' to find the records *similar* to the 'example'.
The focus in this paper will lie on multi-valued attributes (single-valued attributes are special cases of this). As an illustration, consider the instrumental composition of the audio pieces in the archive. Users might be interested in searching for pieces with a similar instrumental composition to that of the piece they have found already.
To allow flexibility in the search, multi-valued attributes coupled with a thesaurus which stores also similarities between different thesaurus terms, will be considered.

In the following Section, the theoretical aspects of calculating the similarity between different types of sets (*regular* sets, multisets, fuzzy sets and fuzzy multisets) is presented. How to apply the theory when comparing multi-valued attributes is subsequently discussed in Section 3. Finally, in Section 4, some conclusions and plans for future work are stated.

2 Similarity in Set Theory

The similarity between two concepts A and B (notation: $sim(A, B)$) is the degree to which A corresponds to B. In general, a similarity relationship needs to satisfy following properties:

- **Reflexivity** $\forall x : sim(x, x) = 1$
- **Symmetry** $\forall x, y : sim(x, y) = sim(y, x)$
- **Transitivity** $\forall x, z : sim(x, z) \geq \sup_y \ \min(sim(x, y), sim(y, z))$

As similarity measure, the Jaccard similarity measure [4] is taken here. The proposal of usage of the fuzzy Jaccard coefficient as a measure of similarity for building fuzzy thesauri is due to Miyamoto [9, 12, 15] who first used it in the context of information retrieval and clustering.
The Jaccard similarity measure is defined as the number of elements shared by the two concepts, divided by the total number of unique elements in both concepts combined. In set theory this can be expressed as the ratio of the number of corresponding elements in A and B, to the total number of elements in A and B together. In this paper only discrete and finite sets will be considered.

[2] In the audiovisual world, meta-data is the contextual information about the music. In the database world, meta-data refers to information about the database and its schema.