LeSiM: A Novel Lexical Similarity Measure Technique for Multimedia Information Retrieval

Ioannis Karydis[1], Andreas Kanavos[2], Spyros Sioutas[1,2], Markos Avlonitis[1], and Nikos Karacapilidis[3]

[1]Dept. of Informatics, Ionian University, Greece
[2]Dept. of Computer Engineering & Informatics, University of Patras, Greece
[3]Dept. of Mechanical Engineering and Aeronautics, University of Patras, Greece
karydis@ionio.gr, kanavos@ceid.upatras.gr, sioutas@ionio.gr, avlon@ionio.gr, karacap@upatras.gr

Abstract. Metadata-based similarity measurement is far from obsolete in our days, despite research's focus on content and context. It allows for aggregating information from textual references, measuring similarity when content is not available, traditional keyword search in search engines, merging results in meta-search engines and many more research and industry interesting activities. Existing similarity measures do not take into consideration neither the unique nature of multimedia's metadata nor the requirements of metadata-based information retrieval of multimedia. This work proposes a customised for the commonly available author-title multimedia metadata hybrid similarity measure that is shown through experimentation to be significantly more effective than baseline measures.

Keywords: Multimedia; Metadata; Similarity measure; Lexical similarity; Author-title.

1 Introduction

Multimedia Information Retrieval (MIR), such as videos, musical content, animation etc, is ubiquitous nowadays. Search engines identify video content pertaining to a query and present results as ready to be consumed in their appropriate mode[1] while musical content providers mine preferences through social networks and other sources in order to assist implicit musical queries leading to playlists[2].

Research related to MIR has long focused on multimedia's content for producing representations on which to perform retrieval related tasks such as similarity measurement. The advancement and widespread penetration of virtual social networks has provided another source of information that is contextual to the actual content and mostly refers to social networks users' interaction with or

[1] http://www.google.com, http://www.bing.com
[2] http://www.last.fm, http://www.spotify.com

related to the multimedia content. Contextual representations have been shown to significantly boost information retrieval related results in an array of scenarios (Melucci, 2008; Karydis, Kermanidis, Sioutas, & Iliadis, 2013).

Despite the aforementioned focus on the content and context derived descriptors from multimedia data, metadata also allow for direct interpretation of their respective multimedia content (Hanjalic, Lienhart, Ma, & Smith, 2008). Metadata descriptors, for all their shortcomings, when existing and accurate, offer a set of mostly predefined textual descriptors that allow for fast and relatively computational cheap information retrieval. Moreover, existing text information retrieval methods can be used, up to a degree of success with almost no adaptation, alleviating thus the need for customised methods for preliminary results.

Numerous approaches as to the schemata that best describe multimedia content exist (Smith & Schirling, 2006). In almost all approaches, the notion of a very short textual description (title) of the content as well as attribution of the content to its author/performer is common a phenomenon. These two attributes, the title and author, although not of the best discriminative capacity exhibit adequate representative capability and are assigned to the content, by its author, *ad hoc*.

Given multimedia's content- and context- based MIR research results, the focus to metadata proposed herein may initially sound anachronistic. Nevertheless, this is far from the truth, as metadata-based MIR is still required for a plethora of research and industry related activities, such as: aggregating multimedia information from textual references (e.g. screen-scraping from html pages), measuring similarity when content is not available (e.g. client-side playlist editing without need to stream data or burden the server), traditional keyword search in search engines, merging results in meta-search engines that do not host content due to intellectual property issues, and many more.

Existing methodologies focusing on metadata-based MIR can be broadly separated into two classes based on whether use of supportive to the actual metadata information is used or not. This supportive information is similar to the aforementioned contextual (but not necessarily from social media), requires collection and although it may enhance the metadata it can also introduce noise (Metzler, Dumais, & Meek, 2007). In this work we focus on solely the title and author metadata information of the multimedia content (i.e. no use of supportive/contextual information is done) in order to perform similarity measurement.

1.1 Motivation & Contribution

Existing methodologies for similarity measurement using solely the metadata of the multimedia content are generic as to the type of text applied onto. Thus, they take into consideration neither the unique nature of multimedia's metadata nor the requirements of metadata-based information retrieval of multimedia.

Thus, in order to achieve a similarity measurement for multimedia author-title metadata, this work:

- proposes a hybrid lexical similarity measure for the common author-title metadata of multimedia entities,

– conducts experimentation in order to verify the increased effectiveness of the proposed similarity measure in comparison to existing methods.

The remaining of the work is organised as follows: Section 2 presents background information and existing research on metadata and short text similarity. Next, Section 3 presents the proposed lexical similarity measure for multimedia author-title metadata. Section 4 details the experimental evaluation of the proposed method, while the work is concluded in Section 5.

2 Background & Related Research

One of the key processes in textual similarity measurement is the representation of the text used in order to apply methods and techniques. Three usually assumed categories of representation (Metzler et al., 2007) are the surface, the stemmed and the expanded. The surface refers to the unaltered text itself while the stemmed is a normalisation / generalisation of the text wherein words are reverted to their stems aiming at the removal of the commoner morphological and inflectional endings (Porter, 1980). Finally, the expanded representation utilises external resources in order to enrich the surface/stemmed representation with contextual information. In this work, we focus on the surface presentation due to the requirements of the problem addressed herein.

2.1 Related Research

Metzler et al. (Metzler et al., 2007) presented a number of similarity measures for short segments of text. Although their work revolves around overcoming the vocabulary mismatch and contextual information problem, and thus is outside the scope of our work, they also propose a hybrid similarity measure that utilises the surface representation as well.

Bearing in mind the shortness of the author-title metadata in terms of number of characters, one may assume that short text semantic similarity and sentence similarity methods lend themselves as applicable approaches in order to tackle the problem addressed herein.

Short Text Semantic Similarity (STSM), in contrast to traditional text similarity methods, such as tf-idf cosine-similarity, aims at semantic level matching (Boom, Canneyt, Bohez, Demeester, & Dhoedt, 2015). The focus on the semantic level of the short texts' similarity is, apart from a required feature to judge similarity of meaning, also due to the lack of word overlap in the short texts compared since these are, more often than not, free text expressions of humans. In contrast multimedia author-title metadata require exact match for the author part (e.g. "The Doors" group have nothing to do with actual doors and no relation to an imaginary author artist titled "The gates" that feature the same notion) and a degree of flexibility for the title part (e.g. "Episode V: The Empire Strikes Back" should be a correct result in either original and remastered versions, usually described in the title in contrast to an episode of the TV-series

titled "The Empire Builds Back"). Moreover, the common size of short texts is far more lengthy (10-20 words (O'Shea, Bandar, Crockett, & McLean, 2008)) than the usual length of approx. 7 words of the concatenated author-title (see Section 4.1). Accordingly, STSM methods are not applicable to the problem tackled herein.

In the same manner, sentence similarity methods are also not applicable to the problem tackled herein as their main focus is on sentences' meaning, usually for the purposes of text summarization and machine translation (Li, Hu, Hu, Wang, & Zhou, 2009). In contrast, multimedia author-title metadata are clearly not selected/assigned to multimedia content on the grounds of syntactical and/or grammar rules, nor for their meaning conveying capabilities.

2.2 Baseline Similarity Measures

In order to compare the performance of the proposed similarity measure, in this Section we formally describe a set of baseline similarity measures. All of these are applied on the surface representation and are invariably lexical, i.e. are matching words / terms between the query q and candidate text c.

Exact In this case q is character per character equal to c, i.e. $\sum_{i=1}^{N} diff(q_i, c_i) = 0$ where both q and c are of length N and the function $diff$ is any symmetric distance measurement function for characters. Due to its nature, this similarity measure returns a binary result, indicating whether or not q is an exact match of c.

Substring This measure relaxes the requirement of holistic exact similarity between q and c, by allowing a match to be established if q is a continuous exact substring of c, i.e. c is a match for q when $\sum_{i=1}^{N} diff(q_i, c_j) = 0$ where $diff$ is as previously defined, q is of length N, c is of length M with $N \leq M$, $1 \leq j \leq M$ and $j = i + \alpha$ where $0 \leq \alpha \leq M - 1$. Accordingly, the substring similarity measurement is evidently a generalisation of the exact similarity measurement with $M = N$ and $\alpha = 0$.

It is obvious that the sizes in number of characters of q and c, $|q|$ and $|c|$ respectively, greatly affect the possibility of identifying c as a match for q, especially when $|q|$ is very small in relation to $|c|$. Thus, if q and c only share a single character substring (e.g. $|q| = 1$), then can labeled a match, a result that intuitively hampers the measurement's performance. Accordingly, in this work we introduce a secondary condition for c to be a match for q, and that is the relative length of q to c that is left as a variable for experimentation purposes.

Subset In this case, q and c are split into terms and each term of q is searched for exact matches with terms of c. In detail, c is a match for q when $q_{terms} \subset c_{terms}$ where q_{terms}, c_{terms} is the set of terms for q and c respectively and the similarity between terms is based on the aforementioned exact similarity measure. Following the same pattern, the subset similarity measure is a generalisation of the substring similarity measure where the requirement for continuity of the substring's terms is alleviated.

As with the case of the substring similarity measure, q and c may only share a single term to be labeled a match (e.g. $|q_{terms}| = 1$ while $|c_{terms}| \gg 1$), again a case that affects the measurement's performance. Thus, we introduce a secondary condition for c to be a match for q, and that is the ratio of $|q_{terms}|$ to $|c_{terms}|$ that is left as a variable for experimentation purposes.

3 Proposed Method

The proposed methodology is based on the subset similarity measure presented in 2.2 but with a number of alterations.

Symmetric similarity Initially, the measure is turned into a symmetric with the calculation of not only c being a match for q but also q being a match for c, and then the ratio of matched to total terms $\frac{|q_{terms}^{matched}|}{|q_{terms}^{total}|}$ and $\frac{|c_{terms}^{matched}|}{|c_{terms}^{total}|}$ are combined using an equally weighted average. This process aims at taking into consideration not only the exact matching subset between q and c but also their relative subset sizes in a manner that is less strict than the ratio of $|q_{terms}|$ to $|c_{terms}|$. Most importantly though, it turns the measure's output from binary to range and can be thus addressed similarly as with the secondary conditions for substring and subset measures.

Synonymy detection In order to address the special characteristics of each multimedia type, a simplistic and targeted notion of synonymy handling is introduced. As the surface representation utilised herein follows the current naming practices of each multimedia domain's, extraneous information is occasionally included in the author-title metadata. A common example is the variable expression for indication of a guest performer in musical content with (a plethora of alternatives of) the "featuring" term or the inclusion of encoding - quality descriptors (e.g. "1080p", "H264", "DD5.1", etc) in videos' titles. Accordingly, for each q's term included in a synonym set, the rest of the synonyms of this set are additionally searched in c without affecting the $|q_{terms}^{total}|$. As such synonymy definition is outside the scope of this work, the proposed methodology is ignorant to its origin that could be based on custom pre-definition or even be learned using machine learning techniques. It should be noted that this notion of synonymy is far from the generic semantic similarity described in Section 2.1 mostly due to its targeted application. In other words, our approach herein focuses on selective term-related semantic similarity in order to address common naming practices of multimedia domains while not affecting the rest of the similarity process.

Approximate matching In contrast to subset similarity measure's function for similarity between terms being based on exact matching, our method uses Levenshtein edit distance that identifies the minimum number of operations required to transform the searched term into the candidate (Levenshtein, 1966). The introduction of approximate matching between terms addresses a very important characteristic of the author-title metadata, that of their accuracy level. Metadata are nowadays mostly assigned to multimedia content by content creators, and thus are as accurate as possible. Nevertheless,

existing content's metadata may well originate from the era that such information was mostly user assigned and thus error prone. Moreover, user created/edited content is bound to include user-created metadata, and thus discrepancies, while also, user generated queries are *sui generis* as to their inaccuracy. The degree of Levenshtein distance in order to assume match between terms is herein left as a variable for experimentation purposes.

4 Performance Evaluation

In support of the efficiency of the proposed similarity measure, this section presents a set of experiments that have been performed. A concise description of the experimentation dataset is also given followed by a performance analysis.

4.1 Experimental Setup

The proposed similarity measure was tested on a dataset of musical content's metadata, that despite its focus on one type of multimedia content, is generic enough in order to both maintain the generality of the proposed solution as well as function as a proof of concept.

The dataset comprises of 100 queries collected from mining textual information of currently playing tracks from popular Greek radio stations, playing English pop music. Each query was then submitted to 5 musical information providing search engines' API, Apple[3], Spotify[4], 7digital[5], Last.fm[6] and MusicBrainz[7] collecting at maximum 30 results per query per search engine. 9.015 results where collected as a result of the queries in total, with an average number of 2,1618 terms per author and 5,1138 terms per title. The collected replies' relevance to their originating query was subsequently manually evaluated using a binary classifier.

The experimentation parameters include, for the baseline measures described in Section 2.2, the secondary condition ratios for both substring, subset and proposed measures, while solely for the proposed method, the degree of Levenshtein distance. In that sense the secondary condition level acts as the degree of similarity between q and c.

The evaluation of the algorithms' results is made by means of precision and recall, that was then combined using the F-Measure (Van Rijsbergen, 1979).

[3] `https://affiliate.itunes.apple.com/resources/documentation/itunes-store-web-service-search-api/`
[4] `https://developer.spotify.com/web-api/`
[5] `http://developer.7digital.com/`
[6] `http://www.last.fm/api`
[7] `https://musicbrainz.org/doc/Development/XML_Web_Service/Version_2`

4.2 Experimental Results

The first experiment presents the performance of the proposed similarity measure for varying degree of Levenshtein distance and degree of similarity between q and c.

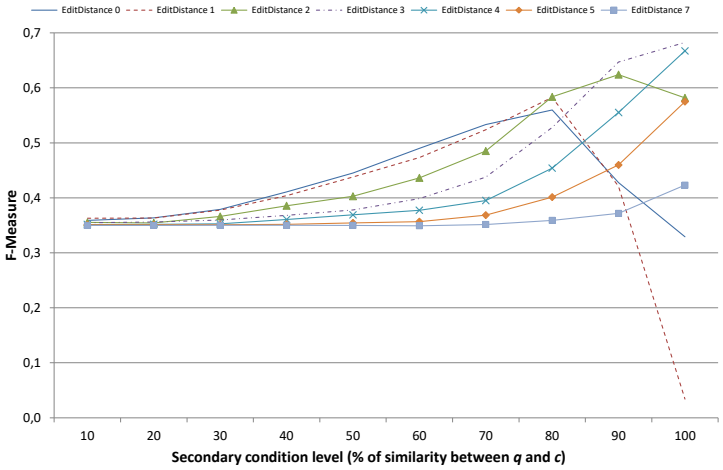

Fig. 1. The performance of the proposed measure for varying degree of Levenshtein distance and degree of similarity between q and c.

Figure 1 shows that the proposed measure is indeed sensitive to the Levenshtein distance, with values of 3 and 4 reaching the best F-Measures, followed by values of 2 & 5 also achieving good performance for the requirement of 100% degree of similarity between q and c. The results indicate that the intuition of incorporating approximation for the terms' matching indeed paid-off, evident by the comparison of performance between Levenshtein distance values 0 and 3. Moreover, the results indicate that approximation for the terms' matching should not be thought of as a panacea, since allowing for more approximation, after a certain point, performance degrades, evident by the comparison of performance between Levenshtein distance values 2, 3 and 7. The abrupt change for Levenshtein distance value 1 is attributed to the relatively small size of the dataset under examination.

The second experiment focuses on the relative performance of the baseline and proposed similarity measures. In this case, the Levenshtein distance for the proposed measure is set to 3 following the best attained result of the previous experiment. Figure 2 shows the attained F-Measure for varying degree of similarity between q and c for all baseline and proposed measures.

The results indicate the superior performance of the proposed methodology especially when the requirement of similarity between q and c is in its strictest setting, that is for the substring method is required for the relative length of q to

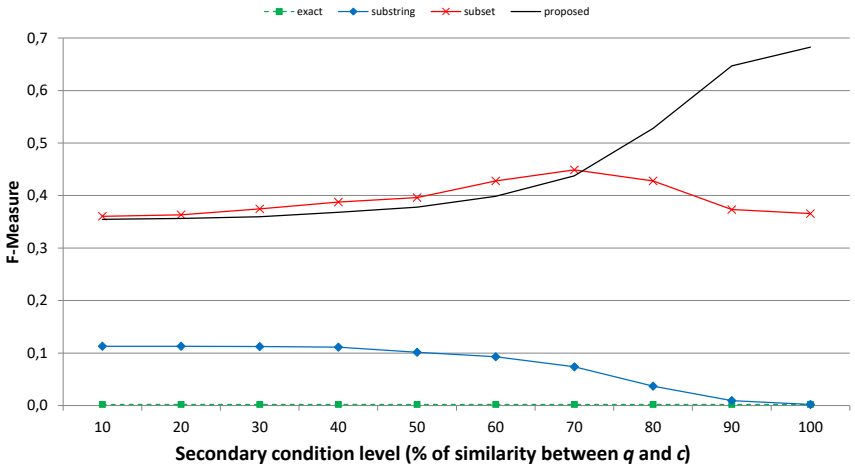

Fig. 2. The performance of all examined similarity measures for varying degree of similarity between q and c.

c to be equal to 1, for the subset the ratio of $|q_{terms}|$ to $|c_{terms}|$ to be equal to 1 and for the proposed the weighted average of $\frac{|q_{terms}^{matched}|}{|q_{terms}^{total}|}$ and $\frac{|c_{terms}^{matched}|}{|c_{terms}^{total}|}$ to be equal to 1. It should be noted that the performance of the exact similarity measure is constant as it is not affected by the secondary condition levels.

5 Conclusion

This work examines the use of metadata-based similarity measurement for the purposes of multimedia similarity measurement. Such measurements as usually done on the content or context space with significant accuracy, though a number of tasks are better off measuring multimedia entities' similarity using solely metadata, and especially, the commonly found author-title information. Such tasks include aggregating information from textual references, measuring similarity when content is not available, traditional keyword search in search engines, merging results in meta-search engines and many more research and industry interesting activities. Existing similarity measures do not take into consideration neither the unique nature of multimedia's metadata nor the requirements of metadata-based information retrieval of multimedia.

Accordingly, herein we propose a customised for the commonly available author-title multimedia metadata hybrid similarity measure. The proposed measure draws from the subset similarity measure with significant alterations such as its conversion to symmetric measure, the targeted inclusion of semantic support for specific sets of search terms and the substitution of the exact matching applied during term similarity with approximate matching.

Results indicate the superiority of the proposed similarity measure in comparison to baseline approaches, as well as the advantage provided by the proposed customisations/alterations.

Future work includes the expansion of the dataset's size and multimedia content type in order to achieve increased generalisation as well as the re-evaluation of the relevance of results obtained for each query using a likert scale increasing thus the quantisation of the evaluation in order to better map the notion of author-title similarity.

References

Boom, C. D., Canneyt, S. V., Bohez, S., Demeester, T., & Dhoedt, B. (2015). Learning semantic similarity for very short texts. In *2015 ieee international conference on data mining workshop (icdmw)* (p. 1229-1234). Retrieved from http://dx.doi.org/10.1109/ICDMW.2015.86 doi: 10.1109/ICDMW.2015.86

Hanjalic, A., Lienhart, R., Ma, W.-Y., & Smith, J. R. (2008). The holy grail of multimedia information retrieval: So close or yet so far away? *Proceedings of the IEEE*, *96*(4), 541–547. Retrieved from http://dx.doi.org/10.1109/jproc.2008.916338

Karydis, I., Kermanidis, K. L., Sioutas, S., & Iliadis, L. (2013). Comparing content and context based similarity for musical data. *Neurocomputing*, *107*(0), 69 - 76. Retrieved from http://www.sciencedirect.com/science/article/pii/S092523121200759X doi: 10.1016/j.neucom.2012.05.033

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady* (Vol. 10, pp. 707–710). Retrieved from http://adsabs.harvard.edu/abs/1966SPhD...10..707L

Li, L., Hu, X., Hu, B. Y., Wang, J., & Zhou, Y. M. (2009). Measuring sentence similarity from different aspects. In *2009 international conference on machine learning and cybernetics* (Vol. 4, p. 2244-2249). Retrieved from http://dx.doi.org/10.1109/ICMLC.2009.5212182 doi: 10.1109/ICMLC.2009.5212182

Melucci, M. (2008, June). A basis for information retrieval in context. *ACM Transactions on Information Systems*, *26*(3), 14:1–14:41. Retrieved from http://doi.acm.org/10.1145/1361684.1361687 doi: 10.1145/1361684.1361687

Metzler, D., Dumais, S., & Meek, C. (2007). Similarity measures for short segments of text. In G. Amati, C. Carpineto, & G. Romano (Eds.), *Advances in information retrieval: 29th european conference on ir research, ecir 2007, rome, italy, april 2-5, 2007. proceedings* (pp. 16–27). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-540-71496-5_5 doi: 10.1007/978-3-540-71496-5_5

O'Shea, J., Bandar, Z., Crockett, K., & McLean, D. (2008). A comparative study of two short text semantic similarity measures. In N. T. Nguyen,

G. S. Jo, R. J. Howlett, & L. C. Jain (Eds.), *Agent and multi-agent systems: Technologies and applications: Second kes international symposium, kes-amsta 2008, incheon, korea, march 26-28, 2008. proceedings* (pp. 172–181). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from http://dx.doi.org/10.1007/978-3-540-78582-8_18 doi: 10.1007/978-3-540-78582-8_18

Porter, M. (1980). An algorithm for suffix stripping. *Program*, *14*(3), 130-137. Retrieved from http://dx.doi.org/10.1108/eb046814 doi: 10.1108/eb046814

Smith, J. R., & Schirling, P. (2006). Metadata standards roundup. *IEEE MultiMedia*, *13*(2), 84-88. Retrieved from http://dx.doi.org/10.1109/MMUL.2006.34 doi: 10.1109/MMUL.2006.34

Van Rijsbergen, C. J. (1979). *Information retrieval*. Butterworth. Retrieved from http://dx.doi.org/10.1002/asi.4630300621