A METHOD FOR EXTRACTING SEMANTIC INFORMATION FROM ON-LINE ART MUSIC DISCUSSION FORUMS

Mohamed Sordo[1], Joan Serrà[2], Gopala K. Koduri[1] and Xavier Serra[1]
[1] Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain.
[2] Artificial Intelligence Research Institute (IIIA-CSIC), Bellaterra, Barcelona, Spain.
{mohamed.sordo,gopala.koduri,xavier.serra}@upf.edu, jserra@iiia.csic.es

ABSTRACT

In this paper a method for extracting semantic information from online music discussion forums is proposed. The semantic relations are inferred from the co-occurrence of musical concepts in forum posts, using network analysis. The method starts by defining a dictionary of common music terms in an art music tradition. Then, it creates a complex network representation of the online forum by matching such dictionary against the forum posts. Once the complex network is built we can study different network measures, including node relevance, node co-occurrence and term relations via semantically connecting words. Moreover, we can detect communities of concepts inside the forum posts. The rationale is that some music terms are more related to each other than to other terms. All in all, this methodology allows us to obtain meaningful and relevant information from forum discussions.

1. INTRODUCTION

Understanding music requires an understanding of how listeners perceive music, how they consume it or enjoy it, and how they share their tastes among other people. The online interaction among users results in the emergence of online communities. These interactions generate digital content that is very valuable for the study of many topics, in our case for the study of music. According to [1], an online community can be defined as a persistent group of users of an online social media platform with shared goals, a specific organizational structure, community rituals, strong interactions and a common vocabulary.

In this paper we propose a method for extracting semantic information from online art-music discussion forums. The method starts by defining a dictionary of standard and culture-specific music terms, and then creates a complex network representation of the online forum by matching such dictionary against the forum posts. The resulting network can then be analyzed using different network measures, including link structure, node relevance, node co-occurrence and term relations via semantically connecting words. This allows us to obtain meaningful information from the forum's discussions.

The rest of the paper is organized as follows. Section 2 provides a brief overview on the state of the art in both information extraction (from on-line discussion forums), and graph based algorithms for Information Retrieval and Natural Language Processing. The methodology for creating a complex network representation of a forum text content is described in Section 3. In Section 4 we present several network measures and discuss their application in extracting relevant information from forum posts. Finally, we draw some conclusions and point out future work in Section 5.

2. BACKGROUND

A considerable number of approaches devoted to mining user-generated text content (such as blogs, reviews, social tags) have been proposed in the music information retrieval (MIR) community (e.g. [2–5]). Nevertheless, to the best of our knowledge, none of these highlighted approaches has exploited the inner structure of online discussion forums. Unlike other types of user-generated text content, on-line discussion forums capture the interaction between different users in a more explicit way. Different opinions and point of views over a topic can be provided/established, and reaching a consensus among all users is not always guaranteed. Hence, extracting information from an online discussion forum could help to reveal relevant aspects of the forum related to user opinions, topic novelties, current tendencies in the field (in this case, art music traditions), etc.

2.1 Information extraction from discussion forums

Extracting semantic information from online forums has become an important area of research (mainly in other fields that take benefit from text processing) in the last few years. For instance, Yang et al. [6] proposed a method to extract structured data from all types of online forums. Weimer et al. [7] and Chen et al. [8] proposed models to identify high quality posts and topics, respectively. Zhu et al. [9], on the other hand, generated relation networks for topic detection and opinion-leader detection.

2.2 Graph-based algorithms for IR and Natural Language Processing

Oftentimes treated as separated research areas, in the last decade there has been a growing interest in using techniques from graph theory and complex networks for information retrieval (IR), and natural language processing

(NLP) [10]. The main idea is to represent textual content, notably web content (such as blogs, news, reviews, forums, etc.) as a graph or network, where the nodes represent single words or a set of words with a particular sense (usually referred to as n-grams) and the edges represent relations between these terms. The resulting network is then analyzed with state of the art complex network measures [11] in order to study its characteristics, or to extract relevant information from it. Research using such graph-based techniques spans a wide range of text processing subjects, including semantic similarity [12, 13], clustering [14], machine learning [15, 16], opinion mining [17], summarization [18, 19], word sense disambiguation [20, 21] or information retrieval [22–25], among others.

3. METHODOLOGY

The proposed method for extracting information from online discussion forums starts by defining a dictionary of culture-specific musical terms (Section 3.1). The content of this culture-specific dictionary can be obtained from existing ontologies and additional resources covering most of the aspects of the studied musical cultures. Once the dictionary is built, the method proceeds by matching such dictionary against the forum posts (Section 3.2). Depending on the posterior analysis of the network, the proposed method can be extended to match additional contextual terms in the forum, such as nouns, adjectives or adverbs. The matched terms are then used to generate a network representation of the forum posts, by assigning a node to each matched term and connecting the nodes with edges if two matched terms are sufficiently close in the text (Section 3.3). Figure 1 shows an example of the network representation of a forum. Once the network is generated, it might be required to filter it, in order to remove irrelevant and noisy information (Section 3.4).

3.1 Dictionary creation

A dictionary of terms is first built to help in identifying and extracting culture-specific music terms from a text. For that purpose, editorial metadata from a music collection, that can be considered as representative of a music repertoire, should be gathered. The metadata could include standard information about music items, such as names of recordings, releases, works (compositions), composers/lyricists and performers, but also information about culture-specific concepts, such as raagas [1] and taalas [2] for Carnatic and Hindustani music [3], or makam and üsul for Makam music in Turkey. This metadata can be extracted easily, for instance, from websites such as MusicBrainz.org, an open music encyclopedia wich aims at storing and providing information related to artists, their works and the relations between them.

The editorial metadata can be extended with additional sources of information, coming from dedicated websites or other encyclopedias of general knowledge. A well known community encyclopedia is Wikipedia. Besides the articles discussing different aspects of a music culture, the community of Wikipedia also provides additional information about categories, which group articles referring to the same subject in a hierarchical form. Following [26], one can obtain a list of culture-specific music terms from dbpedia.org, a machine-readable representation of Wikipedia. We start from a seed category that defines the name of the music culture (e.g., "Carnatic_music", "Hindustani_music") and explore the inherent structure of the dbpedia categorization in order to get all the terms related to the seed. The final dictionary is then created by merging MusicBrainz metadata and Wikipedia categories, and stored as a flat taxonomy of category terms (e.g. raaga–bhairavi, instrument–bağlama, etc.).

The main problem of such a dictionary of terms from art music tradition is that it suffers from noise and spelling errors, mainly due to the diverse transliterations to English of foreign languages terms. For instance, the name *Tyagaraja* (a legendary composer of Carnatic music) can also be written as *Thayagaraja*, *Thiagaraja*, *Tyagayya*, *Thiyagaraja*, *Thagraja*, etc. In order to clean the dictionary, a string matching method based on a linear combination of the longest common sub-sequence and Levenshtein algorithms [27] can be applied to find all duplicate terms, which are further filtered manually in order to maintain a single common description for each of them.

3.2 Text processing

Before building complex network representation of the forum, we apply some text processing techniques to match the generated music dictionary against the forum posts. We iterate over the posts of all the topics of the forum. For each post, the text is tokenized by using any existing tokenizing technique [28] (in our case we use Penn Treebank). The words are then tagged using a part-of-speech (POS) tagger (Maxent Treebank in our case) [28].

Once the text is tokenized and tagged, the method proceeds to match the dictionary of culture-specific music terms against the list of tagged tokens. Given that some terms in the dictionary are word n-grams (i.e. terms with more than one word), the dictionary is sorted in descending order by the number of words, matching the longest terms first. This is done to avoid matching long dictionary n-grams as shorter n-grams or simple unigrams.

In order to capture semantic relationships among musical terms, it might be relevant to add contextual words from the forum posts. Such words can include adjectives, nouns, adverbs, etc. The presence of these words in the forum posts is provided by the POS tagging. Thus, these contextual words are also matched in the forum posts, except for stop words and very short words (i.e., words with fewer than 3 characters).

The unmatched words are not removed from the list of tokens, but rather marked as non-eligible. For example, the sentence "the difference between AbhEri and dEvagAndhAram" is converted to "** difference ** AbhEri ** dEvagAndhAram", where ** denotes a non-eligible word. Al-

[1] Raaga is a fundamental melodic framework for composition and improvisation in Indian classical music.

[2] Taala is a rhythmic framework for composition and improvisation.

[3] The concepts of raaga and taala are the same in Carnatic and Hindustani music but they normally have different spellings.

gorithm 1 summarizes this text processing and dictionary matching step.

Data: $Dict$, a dictionary of music terms ordered by number of words; $Post$, a forum post;
Result: $Terms$, a sequence of terms;
$Terms \leftarrow \emptyset$;
$tokens \leftarrow tokenize(Post)$;
$pos\_tags \leftarrow part\_of\_speech(tokens)$;
$matched\_tokens \leftarrow match\_dict(tokens, Dict)$;
foreach $token \in matched\_tokens$ **do**
  if *(token ∈ Dict) ∨ (token isNoun | pos_tags) ∨ (token isAdjective | pos_tags)* **then**
    $Terms \leftarrow Terms \cup token$;
  else
    $Terms \leftarrow Terms \cup **$;
  end
end

Algorithm 1: Pseudo-code for the text processing of a forum post. The symbol $**$ represents a non-eligible word.

3.3 Network creation

An undirected weighted network is created by iterating over the processed posts. Algorithm 2 describes how a network representation of the forum posts is created. Each matched term is assigned to a node in the network, and an edge/link [4] between two nodes is added if the two terms are close in the text. The link weight accounts then for the number of times two matched terms appear close in the text.

Data: $Terms$, a sequence of terms; L, a link threshold;
Result: $N = (V, E)$, an undirected weighted network with a set of nodes V and a set of edges E ;
$N \begin{cases} V \leftarrow \emptyset \\ E \leftarrow \emptyset \end{cases}$;
foreach $t \in Terms$ **do**
  $V \leftarrow V \cup t$;
  $close\_terms \leftarrow terms\_close\_to\_t\_at\_dist(L)$;
  foreach $close\_t \in close\_terms$ **do**
    $V \leftarrow V \cup close\_t$;
    if *((t, close_t) ∉ E)* **then**
      $E \leftarrow E \cup (t, close\_t, weight = 0)$;
    else
      $Increment\ weight\ of\ (t, close\_t)\ by\ 1$;
    end
  end
end

Algorithm 2: Pseudo-code for the network creation.

Text closeness is defined as the number of intermediate words between two terms. Thus, we introduce a distance parameter (or link threshold) L that will determine which terms are associated with each other. Keeping the unmatched words in the posts (although they are not finally eligible) is important for calculating this distance. Using the example from Step 1, *AbhEri* and *dEvagAndhAram* are considered to be at a distance of $L = 2$. Our assumption here is that words that are closer in text are more likely to be related.

3.4 Network cleaning

Depending on the characteristics of the resulting network from step 2, this step 3 can be followed or skipped. A high ratio of links to nodes — commonly referred to as high average degree — will produce a very dense network, and extracting relevant information from this network will be highly difficult. For instance, networks obtained in previous work [29] contained $24,420$ nodes and $1,564,893$ links, which means an average degree of 128.16, a very high value for such a small network. In addition, we found that the network contained a lot of noise. Many words (especially rare words or misspellings) appear very few times. We therefore introduce another filter, called frequency threshold F, which filters out the nouns and adjectives that appear fewer than F times.

Thresholds L and F yield a more sparse network. However, it could still be possible that some non-statistically significant term relations were reflected in the network links. Thus, the next step consists of applying a sensible filter to the network topology, the disparity filter [30]. The disparity filter is a local filter that compares the weights of all links attached to a given node against a null model, keeping only the links that cannot be explained by the null model under a certain confidence level [5] α. This confidence level α can be thought of as a p-value ($p = 1 - \alpha$) assessing the statistical significance of a link.

4. NETWORK ANALYSIS

The resulting network from the methodology described in Section 3 can be analyzed by using various complex network measures. The aim of these measures is to describe some relevant aspects that are inherent in the structure of the network.

4.1 Node related measures

4.1.1 Degree

The degree of a node in the network is computed as the number of edges incident to that node. With this simple measure we can obtain the most popular nodes in the network. In our art music tradition case, for instance, we are interested in finding out which are the most popular or most discussed musical terms.

4.1.2 Centrality

A measure of centrality attempts to infer the importance of a node in the network. In our case, it can be used to discover the most influential musical terms in the network, and consequently in the discussion forums. For instance, in the particular case of Carnatic music, we are interested in knowing which are the most important raagas and taalas, or

[4] In this paper, we use both edge and link indistinctly to refer to the same concept.

[5] The null model assumes that the strength of a given node is homogeneously distributed among all its links.

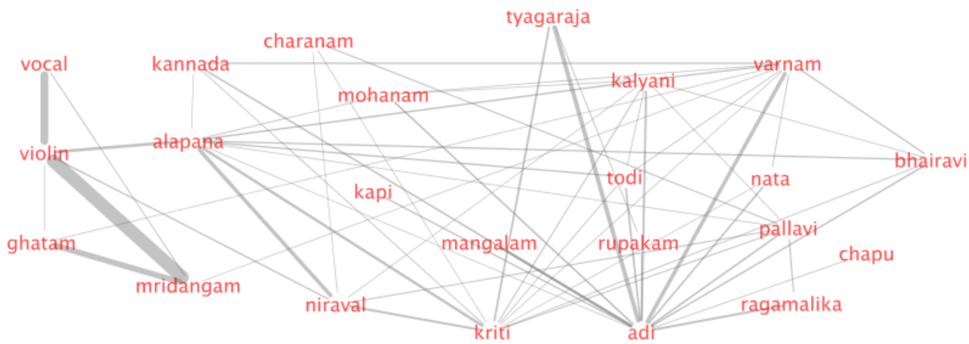

Figure 1. A plot of a subnetwork containing Carnatic music terms with the highest degree. The thickness of the edges represents their weight.

the most influential composers and performers. The same can be applied to other art music traditions.

4.2 Edge related measures

In the proposed network, an edge indicates if two terms co-occur [6] in the same forum post. Thus, an analysis of co-occurrences can reveal the importance of the ties between pairs of terms in the network. There are two possible ways to measure co-occurrence of terms in a network. We distinguish between frequent and relevant co-occurrences.

4.2.1 Frequent co-occurrences

By assuming that terms that co-occur most frequently have a strong relation we can obtain much knowledge from the network. For instance, in [29] we showed how co-occurrent terms allow for correctly guessing the instrument of a performer. In this particular scenario, we rank order the list of instrument neighbors for each performer, based on the weight of the edges, and assume that the highest ordered instrument is more likely to be the instrument of the performer.

4.2.2 Relevant co-occurrences

Although edge frequency already reveals term co-occurrences, it might happen that the weight of some of these edges is not very significant within the network. In that sense, we can also compute a relevance score for the co-occurrence. In the network, this means that we compute a relevance weight for the edge between a pair of nodes. The relevance score $R_{i,j}$ for a link between nodes i and j is obtained by

$$R_{i,j} = \frac{w_{i,j}}{\frac{1}{2}(d_i + d_j)}, \tag{1}$$

where $w_{i,j}$ is the weight of the link and d_x is the degree of node x. This score is giving more relevance to the nodes that are more probable to have some relationship [11, 30].

This relevance measure of co-occurrence can be then applied to combinations of the music term aspects. In [29] we discussed the relation between relevant raaga-raaga and raaga-composer pairs in the case of Carnatic music.

[6] Recall that, in this setup, two terms co-occur if they are at a distance less than L words.

4.3 Network related measures

4.3.1 Community structure

An interesting characteristic that can be measured in complex networks is the discovery of communities. A community in this case can be defined as a set of nodes such that each node is more densely connected to the nodes in that same set than the rest of the network. Although it is hard to extract separable communities, nowadays there exist several methods and approaches that attempt to detect community structure in a network. One such approach is to treat the community structure problem as a clustering problem. Each node is represented as a point in an N-dimensional space, and a similarity distance (e.g., euclidean distance) is computed in order to cluster these points.

In our particular case, community structure can help us, for instance, to discover a strong tie between a certain group of composers and performers, or if a group of composers is more prone to use a particular set of melodic structures more than others. It is interesting to note though that, independent of the method used to extract community structure, the main problem is to interpret these communities, especially when the nodes in the network refer to multiple aspects.

4.3.2 Network structure

The quality and completeness of a network can be evaluated by comparing the network to networks built from other sources of information. For instance, the proposed network can be compared to a network built from the categorization of Wikipedia articles, or from the music items' relations in MusicBrainz. The evaluation can include some of the previously mentioned network measures (node degree, centrality, etc.) in order to detect similarity or dissimilarity between networks.

4.3.3 Semantic relations

Apart from classical network measures, we are especially interested in extracting semantically meaningful relationships between pairs of music terms. From the network perspective, given a pair of nodes, we want to find a third node that is connected to both nodes, and that corresponds to a semantically meaningful relationship concept. We call this node a connecting word. Examples of connecting words (to identify the relationship between pairs of composers and/or performers) include concepts of lineage or family (mother, father, husband, uncle, etc.), musical influence (guru or disciple), similarity (similar, different), etc.

A straightforward approach is to use the same network as before and match the list of predefined connecting words in the common neighbors of a pair of nodes. However, the global nature of the network does not allow us to capture the connecting words correctly, since a connecting word can be related to any of the two compared terms separately. Thus, another approach has to be considered. A possible solution is to apply the proposed methodology locally. That is, instead of creating a single, global network, the method described in Sec. 3 can be applied for each post text individually. For each generated small network, we identify all the common neighbors of a pair of composers and/or performers that are related to the concepts of lineage and musical influence.

5. CONCLUSIONS

In this paper, we presented a method for extracting musically-meaningful semantic information from online art music discussion forums. The method defines a dictionary of culture-specific music terms, and creates an undirected weighted network by matching such dictionary against the forum posts. A post-processing step is applied to clean the network from irrelevant and noisy information. We then discuss the application of some complex network measures to extract meaningful information from the forum posts in a structured fashion.

There are many avenues for future work. First and foremost, we are interested in improving the structure of the network, so that the posterior network analysis can reveal more accurate information. One of the limitations of the current network representation is the lack of more descriptive relations among musical terms. These relations are built upon the fact that two terms that are close in the text are more likely to be related in some way. Assuming that there actually exists a relation between a pair of terms, the network provides no information about what kind of relation this is. For example, a relation between two performers could refer to a collaboration, a family relation or a discussion between two different performance styles. The current network can reveal which are the most relevant or the most influential nodes (i.e., terms) and edges (i.e., relations, co-occurencess), but it does not have the knowledge that the co-occurrences are in fact discussions or other types of relations. Therefore, a more thorough analysis of the textual content in the forum posts is needed. We plan on using more sophisticated NLP techniques, including word sense disambiguation, semantic similarity or summarization. The latter technique can help to reduce noisy and irrelevant information from the forum posts, prior to building the network. Regarding the forum structure, not all the posts or topics are relevant enough to be added to the network. Therefore, we want to find techniques to impose a confidence value per post, depending on the users' relevance to the forum. Another relevant issue to be tackled is the use of a more complete music vocabulary. For that, the metadata that can be found in MusicBrainz and Wikipedia can be extended with information coming from scientific publications (papers, books) or from dedicated expert websites.

Finally, in order to evaluate the generality of the proposed method (i.e., representing a discussion forum as a network of terms and relations using a specific dictionary of terms), we are planning to apply this method in online discussion forums related to different topics, such as films, cars or cooking, among others.

Acknowledgments

This research was partly funded by the European Research Council under the European Union's Seventh Framework Program, as part of the CompMusic project (ERC grant agreement 267583). Joan Serrà acknowledges 2009-SGR-1434 from Generalitat de Catalunya and JAE-DOC069/2010 from Consejo Superior de Investigaciones Científicas.

6. REFERENCES

[1] K. Stanoevska-Slabeva, "Toward a community-oriented design of internet platforms," *International Journal of Electronic Commerce*, vol. 6, no. 3, pp. 71–95, 2002.

[2] M. Schedl and T. Pohle, "Enlightening the Sun: A User Interface to Explore Music Artists via Multimedia Content," *Multimedia Tools and Applications: Special Issue on Semantic and Digital Media Technologies*, vol. 49, no. 1, pp. 101–118, 2010.

[3] B. Whitman and S. Lawrence, "Inferring Descriptions and Similarity for Music from Community Metadata," in *Proceedings of the International Computer Music Conference*, 2002.

[4] P. Lamere, "Social tagging and Music Information Retrieval," *Journal of New Music Research*, vol. 37, no. 2, pp. 101–114, 2008.

[5] O. Celma, P. Cano, and P. Herrera, "Search Sounds: An audio crawler focused on weblogs," in *Proceedings of 7th International Conference on Music Information Retrieval*, Victoria, Canada, 2006.

[6] J. Yang, R. Cai, Y. Wang, J. Zhu, L. Zhang, and W. Ma, "Incorporating Site-Level Knowledge to Extract Structured Data from Web Forums," in *Proceedings of the 18th International Conference on World Wide Web*, 2009.

[7] M. Weimer and I. Gurevych, "Predicting the perceived quality of web forum posts," in *Proceedings of the 2007 Conference on Recent Advances in Natural Language Processing*, 2007.

[8] Y. Chen, X. Cheng, and Y. Huang, "A wavelet-based model to recognize high-quality topics on web forum," in *Web Intelligence and Intelligent Agent Technology, 2008. IEEE/WIC/ACM International Conference on*, 2008.

[9] T. Zhu, B. Wu, and B. Wang, "Extracting relational network from the online forums: Methods and applications," in *Emergency Management and Management Sciences, IEEE International Conference on*, 2010.

[10] R. Mihalcea and D. Radev, *Graph-based natural language processing and information retrieval*. Cambridge University Press, 2011.

[11] M. Newman, *Networks: An Introduction*, M. E. J. Newman, Ed. Oxford University Press, 2010. [Online]. Available: http://www.amazon.com/Networks-Introduction-Mark-Newman/dp/0199206651

[12] A. Budanitsky and G. Hirst, "Evaluating wordnet-based measures of lexical semantic relatedness," *Computational Linguistics*, vol. 32, no. 1, pp. 13–47, 2006.

[13] T. Hughes and D. Ramage, "Lexical semantic relatedness with random graph walks," in *Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007, pp. 581–589.

[14] Z. Chen and H. Ji, "Graph-based clustering for computational linguistics: A survey," in *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 1–9.

[15] A. Blum and S. Chawla, "Learning from labeled and unlabeled data using graph mincuts," in *18th international Conference on Machine Learning*, 2001.

[16] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using gaussian fields and harmonic functions," in *20th international Conference on Machine Learning*, 2003.

[17] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[18] G. Erkan and D. Radev, "Lexrank: Graph-based lexical centrality as salience in text summarization," *Journal of Artificial Intelligence Research*, vol. 22, pp. 457–479, 2004.

[19] R. Mihalcea and P. Tarau, "Textrank: Bringing order into texts," in *Conference on Empirical Methods in Natural Language Processing*, vol. 4, no. 4, 2004.

[20] R. Mihalcea, "Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 2005, pp. 411–418.

[21] E. Agirre and A. Soroa, "Personalizing pagerank for word sense disambiguation," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009, pp. 33–41.

[22] R. Baeza-Yates and B. Ribeiro-Neto, *Modern information retrieval*. Addison-Wesley New York, 1999, vol. 82.

[23] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web." 1999.

[24] J. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.

[25] T. Haveliwala, "Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 15, no. 4, pp. 784–796, 2003.

[26] M. Sordo, F. Gouyon, and L. Sarmento, "A Method for Obtaining Semantic Facets of Music Tags," in *1st Workshop On Music Recommendation And Discovery, ACM RecSys, 2010*.

[27] D. Gusfield, *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press, 1997.

[28] C. Manning and H. Schütze, *Foundations of statistical natural language processing*. MIT Press, 1999.

[29] M. Sordo, J. Serrà, G. Koduri, and X. Serra, "Extracting semantic information from an online Carnatic music forum," in *13th International Society for Music Information Retrieval Conference*, 2012.

[30] M. Serrano, M. Boguñá, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks," *Proceedings of the National Academy of Sciences of the USA*, 2009.