EXTRACTING SEMANTIC INFORMATION FROM AN ONLINE CARNATIC MUSIC FORUM

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

By mining user-generated text content we can obtain music-related information that could not otherwise be extracted from audio signals or symbolic score representations. In this paper we propose a methodology for extracting music-related semantic information from an online discussion forum, rasikas.org, dedicated to the Carnatic music tradition. We first define a dictionary of relevant terms within categories such as raagas, taalas, performers, composers, and instruments, and create a complex network representation by matching such dictionary against the forum posts. This network representation is used to identify popular terms within the forum, as well as relevant co-occurrences and semantic relationships. This way, for instance, we are able to learn the instrument played by a performer with 95% accuracy, to discover the confusion between two raagas with different naming conventions, or to infer semantic relationships regarding lineage or musical influence. This contribution is a first step towards the automatic creation of ontologies for specific musical cultures.

1. INTRODUCTION

Understanding music requires (also) understanding 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 [11], 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. Our aim in this paper is to study and analyze an online community dedicated to the Carnatic music tradition, rasikas.org.

Carnatic music is the art music of south India [12]. This is a very old and alive tradition with a very engaged and active community. The music lovers of Carnatic music are known as rasikas [1], and their involvement in music related activities and events is fundamental for the preservation and evolution of this music. Interestingly, the interactions between artists and rasikas can influence the evolution of the music concepts in the tradition. For instance, raagas [2], often described as collections of phrases, evolve over time (hence, it is often said that a given raaga today is not the same as it was a hundred years ago). When a performer experiments with a new phrase, rasikas respond to show their appreciation if they believe that the phrase enriched their experience of the raaga.

Websites and online forums have become very relevant venues with which to support and sustain the Carnatic music tradition. Online communities have emerged in which groups or rasikas share music content and discuss among them. Rasikas.org is one such forum in which users engage in many types of discussions, some of them quite engaged, covering most relevant Carnatic music related topics. It clearly does not reflect the whole community of rasikas, but it is an interesting forum from which to learn about Carnatic music and about the opinions of some very passionate and active rasikas. We will be using rasikas.org to perform the experiments in this paper.

Extracting semantic information from online forums has become an important area of research in the last few years. For instance, Yang et al. [15] proposed a method to extract structured data from all types of online forums. Weimer et al. [13] and Chen et al. [2] proposed models to identify high quality posts and topics, respectively. Zhu et al. [16], on the other hand, generated relation networks for topic detection and opinion-leader detection. In addition, a considerable number of approaches devoted to mining user-generated text content have been proposed in the music information retrieval (MIR) community (e.g. [1, 4, 8, 14]). Nevertheless, to the best of our knowledge, none of the highlighted approaches in MIR has exploited the inner structure of online discussion forums.

In this paper we propose a method for extracting semantic information from an online Carnatic music forum, specifically rasikas.org. We define a dictionary of Carnatic music terms and create a complex network representation of the online forum by matching such dictionary against

[1] Rasika, in sanskrit, literally means "the one who derives pleasure".
[2] A raaga is the fundamental melodic framework for composition and improvisation in Indian classical music.

the forum posts. We study different network measures related to the aforementioned network, including 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 presents the studied online forum, rasikas.org, and the generated Carnatic music dictionary. The methodology for creating a complex network representation of the forum text content is described in Section 3. We present and discuss results related to the aforementioned network measures in Section 4 and conclude in Section 5.

2. DATA GATHERING

2.1 Dictionary

We first build a dictionary that will help us identify and extract Carnatic music terms from a text. For that we gather the editorial metadata of an extensive list of Carnatic music CD's from MusicBrainz.org, an open music encyclopedia. The metadata includes names of recordings, releases, works (compositions), composers/lyricists and performers, and also information about raagas and taalas [3], two key concepts in Carnatic music. To improve our dictionary we also consider the English Wikipedia as an additional source of information. Similar to [10], we obtain a list of Carnatic music terms from dbpedia.org, a machine-readable representation of Wikipedia. We start from the seed category "Carnatic_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, intrument–mdridangam, etc.).

The main problem of this dictionary is that it suffers from noisiness/misspelling errors, mainly due to the diverse transliterations to English of Indian 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, we apply a string matching method based on a linear combination of the longest common subsequence and Levenshtein algorithms [3] to find all duplicate terms, which we manually filter to maintain a single common description for each of them.

2.2 The rasikas.org forum

The text we analyze is extracted from rasikas.org, a dedicated forum of Carnatic music lovers. As many discussion forums, rasikas.org is divided into different sub-forums, 20 in this case. Each forum contains a list of threads and each thread has a number of posts. Typically, a thread is considered to contain a topic which is discussed in all the posts of that thread. It is interesting to note the ratio of number of posts per thread (Table 1). A median value of 5 means that half of the threads have only 5 posts or less. Regarding

[3] Complementary to raaga, taala is the rhythmic framework for composition and improvisation.

Num. topics	$16,595$
Num. posts	$192,292$
Posts per thread	$\mu = 11.59$, $\sigma = 34.49$, $median = 5$
Num. active threads	1362 active in the last 12 months
Num. users	$4,332$ (with at least one post)
Num. active users	929 active in the last 12 months

Table 1. Some statistics of rasikas.org forum.

the forum users, even though there are 4,332 users in total, only a subset of them, 929, have been active in the last 12 months (a user is considered to be active if she has written at least one post in the last N months and M posts overall).

We crawled the entire rasikas.org forum and stored it locally. Not all the sub-forums are of our interest, though. A few of the sub-forums are not directly related to music, whilst some others discuss topics from other music cultures (e.g. Hindustani). We finally selected a subset of the sub-forums that we considered relevant for our study. This made a total of 11 sub-forums, 14,309 threads and 172,249 posts.

3. METHODOLOGY

3.1 Step 1: Text processing

To extract musically-related semantic information from rasikas.org we generate a complex network representation of it. Nonetheless, before building the network, we apply some text processing techniques to match our Carnatic music dictionary against the forum posts. Hence, in the first step, we iterate over the posts of all the topics for a given subset of sub-forums. For each post, the text is tokenized with Penn Treebank, a classical tokenizing technique. The words are then tagged using the Maxent Treebank part-of-speech (POS) tagger. We use the NLTK toolkit [4] implementation of both the tokenizer and the POS tagger. These methods and implementations are classical choices in natural language processing and computational linguistics [5].

Once the text is tokenized and tagged, the method proceeds to match the dictionary of Carnatic 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 by the number of words, matching the longest terms first. Additionally, stemmed adjectives and nouns provided by the POS tagging are also included, except for stop words and words with less than 3 characters. The non-matched 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.

[4] `http://www.nltk.org`

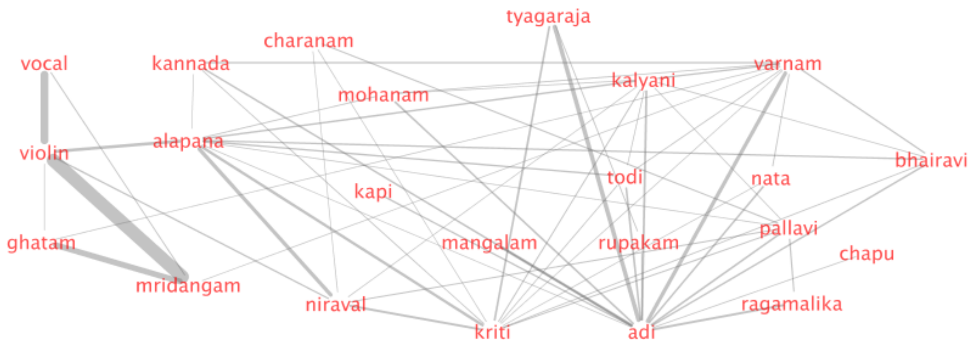

Figure 1. A plot of a subnetwork containing Carnatic music terms with the highest degree. The thickness of the edges represents their weight.

Rank	Raagas	Taalas	Instruments	Performers	Composers
1	Nata	Adi	Violin	Chembai	Tyagaraja
2	Kalyani	Rupakam	Mridangam	Madurai Mani Iyer	Annamacharya
3	Bhairavi	Chapu	Vocal	Charulatha Mani	Purandara Dasa
4	Ragamalika	Jhampa	Ghatam	Kalpakam Swaminathan	Swati Tirunal
5	Kannada	Misram	Morsing	Lalgudi Jayaraman	Papanasam Sivan

Table 2. Top-5 nodes with highest betweenness centrality in the network, organized by category.

3.2 Step 2: Network creation

In the second step, an undirected weighted network is created by iterating over the processed posts. Each matched term is assigned to a node in the network, and an edge/link 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.

Text closeness is defined as the number of intermediate words between two terms. Thus, we introduce a distance parameter L that will determine which terms are associated with each other. Keeping the non-matched 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 is that words that are closer in text are more likely to be related.

3.3 Step 3: Network cleaning

The resulting network from step 2 is very dense (it has $24,420$ nodes and $1,564,893$ links). The average degree is 128.16, which is very high for such a small network. In addition, we find that the network contains a lot of noise. In particular, we find the presence of many spurious nouns and adjectives. Therefore, we introduce a frequency threshold F, which filters out the nouns and adjectives that appear less 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, we decide to apply a sensible filter to the network topology, the disparity filter [9]. 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 α[5]. This confidence level α can be thought of as a p-value ($p = 1 - \alpha$) assessing the statistical significance of a link.

3.4 Network statistics and parameter configuration

After applying the three aforementioned steps, we obtain several network configurations depending on the different parameter values L, F and p. After preliminary assessment, which we omit due to space constraints, we decided to use $L = 5$, $F = 10$ and $p = 0.01$. Fig. 1 shows a subset of the obtained network. With this configuration we obtain a weighted undirected network of $10,928$ nodes (including nouns and adjectives) and $39,067$ edges, resulting in an average degree of 7.15. Furthermore, the network has an average clustering coefficient of 0.309 and a shortest path of 3.658, which are very common values for many other real-world networks [6].

4. RESULTS AND DISCUSSION

4.1 Node betweennes centrality

The resulting weighted network can be analyzed by using different network measures. One of such measures is betweenness centrality. The betweenness centrality measures the importance of a node to the network. It does so by finding the number of shortest paths from all the nodes to all other nodes that pass through that node [6].

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

Parameter configuration	Num. matched performers	Num. matched performer-instrument pairs	Hit %	MRR
$F = 10, L = 5, p = 0.01$	104	63	95.24	95.24
$F = 10, L = 10, p = 0.01$	114	70	80.00	85.48

Table 3. Predicted performer/instrument pairs using frequent co-occurrences.

Table 2 illustrates the top-5 nodes with highest betweenness centrality for each term category. Regarding the raagas, *Kalyani* and *Bhairavi* (two major raagas in Carnatic music) are two frequent choices of artists in order to do a Raagam-Taanam-Pallavi, a complete exposition of their skills in a given raaga [12]. *Ragamalika* is not exactly a raaga per se, but a performance where the performer sings more than a raaga in a single composition. In practice, however, since all the combinations are often named as just *Ragamalika*, without referring to the constituent raagas, the term *Ragamalika* ends up being one of the most frequent ones. *Kannada* is unfortunately an ambiguous term. It can refer to a raaga or to the official language of the south Indian state Karnataka. This is probably why the term might have acquired more weight than, for instance, the raaga *Thodi*, which is in principle more popular. As for taalas, *Adi* is the most preferred taala in Carnatic compositions (511 out of 917 recordings in our Carnatic music collection in MusicBrainz are performed in *Adi* taala). In the case of Carnatic music performers, *Lalgudi Jayaraman* is one of the violin trinity of Carnatic music, the three violinists who are considered the irrefutable masters of the art. Moreover, he is also a renowned composer in the modern times, which can explain why he was ranked high in the betweenness centrality. Finally, regarding Carnatic composers, *Tyagaraja*'s compositions constitute a significant proportion of the Carnatic music repertoire (102 out of the 293 works in our Carnatic music collection were composed by *Tyagaraja*). *Annamacharya* is also a composer with a large number of *keerthanas*, many of which are sung either at the beginning or towards the end of a Carnatic music concert. Keerthana is different from, and generally not as elaborate as, the compositional form called Kriti. Kriti forms the crux of the compositional repertoire in Carnatic music, and is considered to have evolved from Keerthana. *Purandara Dasa* is called the father of Carnatic music, both for his compositions and his systematization of Carnatic music learning process. *Papanasam Sivan* was a prominent composer from the state of Tamilnadu. Indeed, he is often referred to as Tamil Tyagaraja. It happens that many users of rasikas.org are from that region too.

4.2 Term co-occurrences

In this section we analyze co-occurrences of Carnatic music terms. 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. As an example, we will show that network co-occurrent terms allow for correctly guessing the instrument of a performer. This is a non-trivial task whose accuracy can be evaluated objectively, since ground truth data is relatively easy to obtain. In particular, we use the same sources from which we derived the Carnatic music dictionary, Wikipedia and MusicBrainz. From MusicBrainz, we get all the relations between performers and recordings. In the case of Wikipedia, we match a list of instruments (defined in our dictionary) against the text of Wikipedia abstracts. The ground truth is finally filtered by removing the cases where a performer has more than one instrument (vocals are also considered instruments here).

To evaluate the instrument-performer pairing task we use the weight of the links to rank the list of instrument neighbors for each performer (i.e. to rank connected nodes representing an instrument). Two measures are then used to evaluate the predicted instruments in the network: hit percentage and mean reciprocal rank (MRR). The hit percentage refers to correctly predicting the instrument in the first rank. Sometimes, however, the correct answer does not fall in the first rank, but rather in the second, third, or other ranks. This can be measured with MRR, which we define as

$$MRR = \frac{100}{N} \sum_{n=1}^{N} r_n, \qquad (1)$$

where N is the number of co-occurring instruments and r_n is 1 if and only if the n-th instrument is the correct one for a given performer.

As we can observe in Table 3, by considering simple co-occurrences in the network we already achieve an accuracy of 95%, which means that we correctly predict 60 out of the 63 performer-instrument pairs available. For comparative purposes, we also include the results obtained by increasing the link threshold parameter L, which increases the density of the network substantially. Even though the latter configuration increases the number of matched performers and instruments per performers, the accuracy of the predicted instruments drops significantly, meaning that a larger value of L is also adding noise to the network.

4.2.2 Relevant co-occurrences

Apart from evaluating co-occurrences by their frequency, we also compute a relevance score for the co-occurrence.

Raaga	**Raaga**	*Relevance*
Kedaram	Gowla	0.121
Bhavani	Bhavapriya	0.109
Manavati	Manoranjami	0.092
Kalavati	Yagapriya	0.088
Nadanamakriya	Punnagavarali	0.081

Table 4. Relevant co-occurrences of raagas with other raagas.

Raaga	**Composer**	*Relevance*
Abhang	Tukaram	0.159
Yaman kalyani	Vyasa Raya	0.149
Pharaz	Dharmapuri Subbarayar	0.143
Reethi Gowlai	Subbaraya Sastri	0.122
Andolika	Muthu Thandavar	0.108

Table 5. Relevant co-occurences of raagas with composers.

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

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

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 [6, 9].

We apply this relevance measure of co-occurrence to combinations of the previously-mentioned term categories. In this experiment we study two such combinations: raaga-raaga and raaga-composer. Tables 4 and 5 show the top-5 more relevant co-occurrences of these two combinations, which we now comment.

Raaga-raaga By looking back at the rasikas.org discussion forum, we can confirm that the co-occurrences found in Table 4 are related to discussions between different raagas, or similar raagas with different naming conventions. For instance, *Bhavapriya* raaga is also called *Bhavani* by disciples of the famous composer Muthuswami Deekshitar. For the co-occurrence of *Kalavati* and *Ragapriya*, a discussion in a forum thread indicates that *Kalavati* and *Ragapriya* are often confused with each other due to a few historic reasons (e.g. naming convention). Some members of the forum posted several facts and technical arguments that are tied to both raagas [6]. It should be noted that although our method helps us to find these important discussions in the forum, it does not have the knowledge that they are in fact discussions. We believe that extracting contextual information of the co-occurrences in the text could help to predict the type of relation between two terms.

[6] http://www.rasikas.org/forum/viewtopic.php?t=14435

Raaga-composer Intuitively, a relevant co-occurrence of a raaga and a composer might mean that a composer is known by or uses more frequently a particular raaga. Indeed, that is the case of the relations in Table 5. *Vyasa Raya*'s most famous composition, "Krishna Nee Bagane" is in *Yaman Kalyani* raaga. *Dharmapuri Subbarayar* has a popular composition in *Pharaz* raaga called "Smara Sundaranguni". It is also the case of "Sevikka Vendumayya" composition by *Muthu Thandavar* which is in *Andolika* raaga . Contrasting a little bit with these good agreements, the most relevant raaga-composer co-occurrences include *Abhang*, which is a devotional poetry and not a raaga. This is due to a misleading tag in our vocabulary suggesting that *Abhang* is a raaga name, which suggests that a more accurate cleaning process is needed for some terms.

4.3 Term semantic relations

The aim of this last experiment is to extract semantically meaningful relationships between pairs of Carnatic music terms. From the network perspective, given a pair of nodes, we want to find another node that is connected to both nodes, and that corresponds to a semantically meaningful relationship concept. We call this node a connecting word. For this experiment we use a predefined list of connecting words, including concepts of lineage or family (mother, father, husband, uncle, etc.) and musical influence (guru or disciple) to identify the relationship between pairs of composers and/or performers.

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 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. This experiment was evaluated manually using Wikipedia and by asking a Carnatic music expert.

From a total of 24 relations found in the network, our method correctly infers 14 (58% of accuracy). A closer look at the misclassified term relations reveals that many of the wrong predictions were due to ambiguity problems of natural language. For example, our method assigns the relation *guru* for the following pairs of performers: (*Karaikudi Mani*, *G. Harishankar*) and (*Karaikudi Mani*, *Mysore Manjunath*). However, the term *guru* can also be used as a prefix name of a person, in this case *Guru Karaikudi Mani*. Finally, given for instance the following sentence, "Abhisek is grandson of Palghat Raghu and disciple of P.S. Nayaranaswamy", and since the name *Abishek* is not included in our dictionary, our method incorrectly infers a relation of disciple between *Palghat Raghu* and *P.S.*

Nayaranaswamy. It is clear that more advanced natural language processing techniques should be applied in order to solve these ambiguities. Nevertheless, our method is already inferring some relations with a certain amount of confidence, using very simple heuristics.

5. CONCLUSION AND FUTURE WORK

We presented a method for extracting musically-meaningful semantic information from an online discussion forum, dedicated to the Carnatic art music tradition. For that, we defined a dictionary of Carnatic music terms (from concepts such as raagas, taalas, performers, etc.), and created an undirected weighted network by matching such dictionary against the forum posts. Three experiments were ran to study different characteristics of the resulting network, extracting valuable information such as term relevance, term co-occurrence and term relations via semantically connecting words. In the first experiment, nodes with higher betweenness centrality were usually highly correlated with the popularity of Carnatic music terms. In the second experiment, we showed that our method is able to predict the instrument of a performer with a 95% of accuracy. Furthermore, we were able to identify node pairings that are relevant discussions in the forum, using a relevant co-occurrence measure. Even though these discussions were found, our methodology does not have the knowledge that the co-occurrences are in fact discussions. Extracting contextual information of the co-occurrences in the text could help to predict the type of relation between two terms. In the last experiment, we showed that with simple heuristics one can predict semantic relations related to lineage and musical influence with a certain level of confidence.

There are many avenues for future work. Besides the extraction of contextual information and the use of more sophisticated natural language processing techniques, we plan to explore methods that can capture users' opinions using, for instance, algorithms from sentiment analysis [7]. 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 Carnatic music vocabulary. For that, we plan to manually improve the metadata that can be found in MusicBrainz and to use more resources to increment the size of the dictionary. In order to ease reproducibility of our work and to stimulate further research on the topic, the data and code used for these experiments are publicly-available [7].

6. 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 AEDOC069/2010 from Consejo Superior de Investigaciones Científicas, 2009-SGR-1434 from Generalitat de Catalunya, TIN2009-13692-C03-01 from the Spanish Government, and EU Feder Funds.

[7] http://www.dtic.upf.edu/\%7Emsordo/research/ismir2012/index.html

7. REFERENCES

[1] O. Celma, P. Cano, and P. Herrera. Search Sounds: An audio crawler focused on weblogs. In *Proc. of 7th Intl. Conf. on Music Information Retrieval*, Victoria, Canada, 2006.

[2] Y. Chen, X.Q. Cheng, and Y.L. Huang. A wavelet-based model to recognize high-quality topics on web forum. In *Web Intelligence and Intelligent Agent Technology, 2008. IEEE/WIC/ACM Intl. Conf. on*, 2008.

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

[4] P. Lamere. Social tagging and Music Information Retrieval. *Journal of New Music Research*, 37(2):101–114, 2008.

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

[6] M.E.J. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[7] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[8] 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*, 49(1):101–118, 2010.

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

[10] 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.

[11] K. Stanoevska-Slabeva. Toward a community-oriented design of internet platforms. *Intl. Journal of Electronic Commerce*, 6(3):71–95, 2002.

[12] T. Viswanathan and M.H. Allen. *Music in South India*. Oxford University Press, 2004.

[13] M. Weimer and I. Gurevych. Predicting the perceived quality of web forum posts. In *Proc. of the 2007 Conf. on Recent Advances in Natural Language Processing*, 2007.

[14] B. Whitman and S. Lawrence. Inferring Descriptions and Similarity for Music from Community Metadata. In *Proc. of Intl. Computer Music Conf.*, 2002.

[15] 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 *Proc. of the 18th Intl. Conf. on World Wide Web*, 2009.

[16] T. Zhu, B. Wu, and B. Wang. Extracting relational network from the online forums: Methods and applications. In *Emergency Management and Management Sciences, IEEE Intl. Conf. on*, 2010.