Ranking Method Specialized for Content Descriptions of Classical Music

Taku Kuribayashi
Faculty of Engineering,
Kyoto University
Kyoto, Japan
kuribayashi@db.soc.i.kyoto-u.ac.jp

Yasuhito Asano
Graduate School of
Informatics, Kyoto University
Kyoto, Japan
asano@i.kyoto-u.ac.jp

Masatoshi Yoshikawa
Graduate School of
Informatics, Kyoto University
Kyoto, Japan
yoshikawa@i.kyoto-u.ac.jp

ABSTRACT

In this paper, we propose novel ranking methods of effectively finding content descriptions of classical music compositions. In addition to rather naive methods using technical term frequency and latent Dirichlet allocation(LDA), we proposed a novel classification of web pages about classical music and used the characteristics of the classification for our method of search by labeled LDA(L-LDA). The experimental results showed our method performed well at finding content descriptions of classical music compositions.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

vertical search; classical music; labeled LDA

1. INTRODUCTION

When you listen to classical music, you can enhance your understanding of the music by reading content descriptions, especially if you are a non-expert of music, such as an amateur player, because it is difficult to understand the music only from reading the score. A description of the content of classical music is defined as an objective description about the structure of the composition that technically explains specific parts of it. An example content description for Beethoven's 9th Symphony is "The opening theme, played pianissimo over string tremolos, so much resembles the sound of an orchestra tuning." This explains the instruments (strings) and techniques (pianissimo, tremolos).

It is difficult to find useful web pages containing content descriptions of classical music with conventional search engines; if you search by the title of a music, then famous commercial sites such as Amazon or iTunes are highly ranked although they usually do not contain content descriptions. A possible project to help to find such useful pages is a manual collection of links to them; there was such a project in the past [1], while it was suspended because of the amount of work required to maintain the quality. Thus, our objective is to create a ranking that enables you to collect useful descriptions of classical music effectively.

As we surveyed the web pages found by searching with titles of classical music compositions, we discovered that those search results have an interesting characteristic; they can easily be classified into a small number of categories. This characteristic is suitable for applying latent Dirichlet allocation(LDA)[2] and its supervised version, labeled LDA(L-LDA)[3] to the results. Jia et al. [4] applied L-LDA to ranking pages related to a scientific paper. Their method automatically learns keywords assigned to several papers. On the other hand, web pages about classical music are rarely annotated with tags or keywords. Therefore, L-LDA cannot be applied to ranking classical music pages without a proper classification.

In this paper, we propose four methods of re-ranking the web search result, the first two of which are rather naive and the latter two utilize the characteristic explained above: (1)Technical Term Frequency based Ranking(TTFR), (2)LDA based Ranking(LR), (3)L-LDA based Ranking exploiting our classification specialized for web pages about Classical music(LLRC), and (4)LLRC with our additional training data constructed from Wikipedia(LLRCW). The utilized characteristic would be common in search results for traditional cultures such as impressionism paintings and porcelain of the Song Dynasty. Therefore, our idea might be applied to finding useful pages about such cultures.

2. NAIVE METHODS

Two of the methods we propose are rather naive. The first one, TTFR, is based on the hypothesis that good content descriptions contain more technical terms. In TTFR, web pages with higher ratio of technical terms to total number of words in the page are ranked higher. The list of technical terms was created based on "Glossary of musical terminology" and "List of musical instruments" of Wikipedia.

The second one, LR, is based on a hypothesis that pages with good descriptions use similar vocabulary. We select a latent topic corresponding to content descriptions manually from the result of applying LDA to Wikipedia pages of "Symphonies" category. LR ranks pages using the probability of each page belonging to the selected topic. Based on previous researches[5] and preliminary experiments we set the number of topics K as 10, and for the Dirichlet parameters, we used $\alpha = 50/K$ and $\beta = 0.1$.

3. L-LDA BASED RANKING

3.1 Classification for Classical Music

We examined 1540 pages from web search by classical music titles and discovered that they can generally be classified into eight categories. From the observation, we propose a novel way of classification of descriptions on web pages about classical music, which is as follows: (Each page can have more than one label.) **Structure**: Descriptions of the content and structure of a specific part of

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

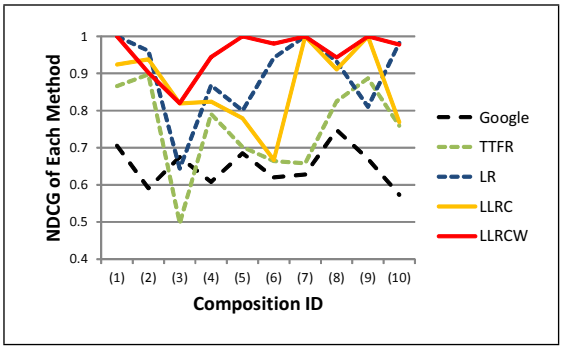

Figure 1: NDCG of our proposed methods and Google ranking

Figure 2: An image of an imaginary application

a composition using objective expressions, such as names of instruments. (104 pages) **Background**: Descriptions of the background of the composition, such as descriptions of the composer and the motive of the composition. (180 pages) **Commentary**: Commentary or evaluation of the composition or a performance of the composition. (169 pages) **Score**: Web pages of sales and downloads of the sheet music. (98 pages) **Cdmp3**: Web pages of CD or mp3 sales, and web pages with only CD track information or videos. (459 pages) **Noneng**: Web pages in languages other than English. (159 pages) **Dictionary**: Web pages of dictionary or encyclopedia articles, only having a simple description. (90 pages) **Irrelevant**: Web pages that do not fit into the labels above. (474 pages)

3.2 LLRC and LLRCW

For our proposed method of LLRC, we used the 1540 pages we explained in the previous section as the training data of L-LDA. Then, we used the distribution of "structure" label as the score of web pages. Based on preliminary experiments, the Dirichlet parameters were set as $\alpha = 50/K$ (K = the number of labels) and $\beta = 0.1$.

There were only 104 pages labeled "structure" in the 1540 pages we used for the method above. LLRCW increases the amount of training data by adding 190 sections of Wikipedia which have high probability, at least 0.3, to belong to the topic used in LR.

4. EXPERIMENTAL EVALUATION

For the experimental evaluation, we listed 10 classical music compositions, and collected top 50 web pages by searching with the title using Google web search for each of them. We eliminated YouTube pages from the list beforehand, because they rarely contain information we aim to obtain.

For each page, we assigned a score on the scale of 0-3, based on how much it contains content descriptions, and we compared the NDCG(Normalized Discounted Cumulative Gain) of the re-ranking methods we proposed.

Figure 1 shows that LLRC and LLRCW improved the NDCG in all of the compositions, and LLRCW showed the best performance on the average, considerably better than naive methods of TTFR and LR.

For example, a good page[1] ranked relatively low(22nd) by Google but 1st by LLRCW includes descriptions such as "The final section, *Nachtwandlerlied*, makes subtle use of tonal and thematic cues," which explains the composition in detail. The structure of the web page is similar to those of Wikipedia, which enabled LLRCW to perform well.

[1]http://whitgunn.freeservers.com/Davemusic/S/strauss-richard/also-sprach-zarathustra.html

5. CONCLUSION

In this paper, we presented methods of effectively acquiring content descriptions of classical music. Our methods, especially LLRCW, improved the NDCG compared to the original Google ranking.

One possible application of our research would be an automatic association system of music(video) and content description system, as shown in Fig.2, using other researches on music analysis, such as [6]. It would be an innovative system to support people who are unfamiliar with classical music enjoy performances more.

6. REFERENCES

[1] Y. Fineman. DW3 Classical Music Resources: Managing Mozart on the Web. *Libraries and the Academy*, 1(4):383–389, 2001.

[2] D. M. Blei, A. Y. Ng, and M. I Jordan. Latent Dirichlet Allocation. *JMLR*, 3:993–1022, 2003.

[3] D. Ramage, D. Hall, R. Nallapati, and C. D Manning. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP 2009, Volume 1*, pages 248–256, 2009.

[4] H. Jia and X. Liu. Scientific Referential Metadata Creation with Information Retrieval and Labeled Topic Modeling. In *iConference 2013*, pages 274–288, 2013.

[5] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(Suppl 1):5228–5235, 2004.

[6] A. Maezawa, M. Goto, and H. G. Okuno. Query-By-Conducting: An interface to retrieve classical-music interpretations by real-time tempo input. In *ISMIR 2010*, pages 477–482, 2010.