SOUND, MUSIC AND TEXTUAL ASSOCIATIONS ON THE WORLD WIDE WEB

Ian Knopke
Music Technology
McGill University
ian.knopke@mail.mcgill.ca

ABSTRACT

Sound files on the World Wide Web are accessed from web pages. To date, this relationship has not been explored extensively in the MIR literature. This paper details a series of experiments designed to measure the similarity between the public text visible on a web page and the linked sound files, the name of which is normally unseen by the user. A collection of web pages was retrieved from the web using a specially-constructed crawler. Sound file information and associated text were parsed from the pages and analyzed for similarity using common IR techniques such as TFIDF cosine measures. The results are intended to be used in the improvement of a web crawler for audio and music, as well as for MIR purposes in general.

1. INTRODUCTION

As has been previously indicated, the "World Wide Web is not a homogeneous, strictly-organized structure" [7]. While much of it may appear random at first glance, there is often an element of similarity in content between many linked web resources. This is in fact a fundamental element of the way in which the entire web functions. It is precisely the embedding of content descriptions within hypertext links that makes web browsing possible, and is one of the major influences behind the popular growth of the Internet.

The main purpose of this paper is to address the relationship between audio and music materials on the web, and the associated web pages that link to them. While this relationship has been studied to some degree between hyperlinked web pages, the structure of audio information on the web remains largely unstudied at the present time. To date, very little attention has been paid to the relationship between web pages and the accompanying audio files. This is surprising, considering the effect of P2P systems and electronic delivery systems on the entire music business. With the success of iTunes and other Internet-based delivery systems, it seems likely that the web will become an important delivery system for music and audio in the near future.

The hypertext/audio relationship is also potentially valuable for Music Information Retrieval (MIR) research, in the sense that web pages with links to sound files form a kind of user-created annotation that is not available in many other MIR contexts. This is potentially an enormous source of information for the MIR field which has not generally been exploited to date, perhaps because of a lack of tools with which to study the problem.

This paper is divided into several sections. After the introduction, the motivations behind the study and how they relate to building a focused web crawler are related, as well as previous and related research. The method used in conducting the experiment is outlined, followed by a discussion of the results. Some directions for future research are also given.

2. MOTIVATION

One of the primary uses of this information is for the creation of a *focused web crawler* [5]. General purpose search engines, such as Google or Lycos, have traditionally tended to use breadth-first crawling techniques, following all links in turn from each web page, then following all links from each child, and so forth. This technique has been shown to effectively find every page, as well as giving good results in situations where all information must be indexed for general purpose queries. Eventually a breadth-first crawl will capture all pages on the web, although this may take considerable time and resources.

However, this may not be desirable in cases where engines are designed to concentrate on a specific topic, class, or type of information. An alternate procedure known as focused web crawling works by prioritizing some links over others and following those links which are most likely to produce the best possible results earlier in the crawling process. This reduces the number of pages that must be parsed to give better quality indexes of pages containing information about a particular topic, and ultimately enhancing the end user experience. This is particularly valuable in the case of a web crawler designed to look for and analyze audio and music files, which require considerable resources beyond what is usually required by an

HTML-only web crawler. Considering the sparseness of audio information on the web, an understanding of the relationship between HTML and the accompanying audio files may considerably reduce the amount of web information which a crawler must accommodate.

Another important consideration is the "freshness" of the information that has been indexed. Not only is the web growing at a rapid rate, but much of the information is changing constantly as web pages are altered or moved to new locations. A web-based search engine is faced with the dual challenge of adding new information, as well as keeping indexes of existing information up to date. A focused crawler, based on a particular topic, can benefit greatly by concentrating on resources that are likely to give the best results. This improves the quality of the query indexes by reducing the number of false nodes and missing pages returned in response to user queries.

3. PREVIOUS RESEARCH

Most large-scale search engines are proprietary commercial ventures that do not make detailed information about their web crawlers public. Few systems are described in detail. Some details of the original Google search engine, as it existed as a Stanford University research project are available [3]. Mercator [11] is a high-capacity web crawler from Compaq Systems Research Center that can be easily configured to handle different types of protocols, such as HTTP, FTP, or GOPHER. Mercator formed the basis of the original AltaVista engine, and has been made available for select research projects. Another recent multithreaded Java-based crawler uses multiple machines connected using the Network File System (NFS) to transfer data between nodes [18]. Other crawler designs have been described in less detail [20, 9, 4, 2]. Crawling technology for multimedia is still in its infancy, with most development in this area applied to still images [16, 17, 10]. The Cuidado project is an audio and music content analysis system that is intended to eventually incorporate a web crawling system [19].

There have been several studies of web structure and the hypertext linkages between web pages. Kleinberg [12] identified two important types of web structures. *Authorities* are pages which contain an extensive amount of high-quality information about a set of related topics. *Hubs* are pages which contain multiple links to authorities on a particular topic. The two structures reinforce one another: good hubs have many links to authorities, and good authorities are linked to by many hubs. The Google PageRank algorithm [15] exploits the ability of users to add new links to favored resources, to produce page rankings based on the popularity of a web page. The technique has proven to be enormously successful, although it requires the capture of a significant portion of the web to work reliably. A similar method, using textual similarities between multiple links to the URL resource was proposed by Amitay [1]. Dean and Henzinger [8] proposed a method for finding web pages based almost-entirely on link structure.

4. EXPERIMENTAL METHOD

A crawl of approximately 2,000,000 unique web pages was conducted using the AROOOGA system, a specialized web crawler and search engine designed to locate and index both sound files and associated pages on the web. AROOOGA stands for Articulated Resource for Obsequious Opinionated Observations into Gathered Audio.

Duplicate URLs were removed during the crawling process, yielding a set of unique pages. Crawling was done using a breadth-first search method, a practice that has been shown to give a fairly natural distribution of web resources [13]. Crawling was initiated from a set of 20 starting seeds representing McGill University, as well as several large portals.

Retrieved pages were specifically analyzed for hyperlinks referencing sound or music files. Each link also contains an associated piece of "anchor" text; this is the visible element on a web page that the user "clicks" on to access the sound file. Additionally, studies have shown that words surrounding a hyperlink often contain descriptive text referring to the hyperlink [7, 6]. In this case, a maximum of ten words preceding and following the link were also captured and analyzed, or as many as could be obtained before the occurrence of another hyperlink.

These three pieces of information, the sound file name, anchor text, and surrounding text are considered for this study to form a single document. All three type of information were then parsed into tokens, lowercased, and stoplisted to remove common terms that have little value for comparison. Parsing sound file names into tokens presents special problems not present in other sources [14], mostly due to contiguous nature of the words in the name. After some experimentation, it was determined that tokens could also be derived by carefully splitting on capitalization boundaries, punctuation, and digits (with some exceptions). File name extensions were removed, but used to determine the format of the sound file (WAVE, MP3, etc.) The extracted information was then used to determine commonality between pages and sound files.

4.1. Similarity Measures

A series of textual similarity measures, adapted from IR, are used here to study the relationship between sound files and the linked web pages, similar to the ones proposed by Davison [7] for the purpose of studying topical locality. Results are averaged across all documents. All measures are comparable in that they produce values between 0 and 1, with 1 indicating identical documents, and 0 representing documents without any common terms.

4.1.1. TFIDF cosine similarity

$$TFIDF_{kd} = \frac{TF_{kd} \times IDF_k}{\sqrt{\sum (TF_{kd} \times IDF_k)^2}}$$

where

$$IDF_k = \log \frac{D_N + 1}{D_k}$$

and

$$TF_{kd} = \log(f_{kd} + 1)$$

where IDF_k is the inverse document frequency of term k, D_N is the total number of documents, D_k is the number of documents containing keyword k, and f_{kd} is the number of occurrences of keyword k in document d. Comparisons are accomplished using a cosine measure.

$$TFIDF{-}COS_{kd} = \frac{\sum_{all} TFIDF_{kd1} \times TFIDF_{kd1}}{sqrt(\sum TFIDF^2_{kd1} \times \sum TFIDF^2_{kd2})}$$

4.2. Term Probability

$$F_k = \frac{f_{kd}}{d_n}$$

and

$$Prob_{d_1,d_2} = \sum F_k$$

where k must be an element of both documents.

4.3. Document Overlap

$$Overlap_{d_1,d_2} = \sum \min(F_{k,d1}, F_{k_d2})$$

5. RESULTS

From the initial crawl, 1,726,054 unique pages were found to contain links to valid pages. These were successfully downloaded and parsed. Of this original set, 4500 pages were found to have references to sound files, producing unique links to 35,481 sound files, a density of less than one percent (0.26%) of pages crawled.

The average number of sound file links per web page across the entire corpus is 0.024. However, the average number of sound file links, on pages which contain references to sound files is 7.71, with the largest number of links on a single page being 341.

The list of sound files is segmented by file format in Table 1. The majority of files (66.33%) were found to be MPEG types, with the next largest category being WAVE files (31.99%).

Type	Number	Percent	Avg Size(Kb)
MPEG	23533	66.33	1662
WAVE	11349	31.99	701
AIFF	578	1.63	2301
OTHER	21	0.06	586
TOTAL	35481	100	1637

Table 1. Sound File Types

5.1. Common Tokens

From the entire set of extracted tokens, lists of the most common terms were computed for each of the three categories of information. By far, the most common token found was "mp3", most often occurring in statements such as "click here to get the mp3". Other terms indicate a strong pop and classical music presence on the web. It is interesting that many of the most common tokens (click, track, download, kb, audio, file) appear to describe elements of the web or file structure, rather than the content of the file.

Sound file names
classical(1604) archives(1602) wav(1304) pre(1304) music(1302) opus(465) track(411) buzz(409) piano(292) demo(282) bach(261) bass(209) love(204) chopin(197) sonata(175) project(167) mozart(167) mix(145) brahms(134) guitar(132)
Anchor/link text
mp3(2168) listen(1435) download(712) demo(494) sample(405) kb(346) mono(324) click(321) mpeg(285) clip(264) wav(260) audio(224) hear(216) guitar(193) sound(184) normal(173) song(156) excerpt(139) bass(125) tone(124)
Text surrounding link
kb(2804) mp3(1997) sound(1529) wav(993) sample(791) code(770) perform(756) file(722) audio(637) drum(613) loop(532) guitar(508) cd(495) record(457) origin(425) song(420) clip(419) listen(396) plain(387) music(386)
Combined
mp3(4259) kb(3165) wav(2557) listen(1845) classical(1818) sound(1778) archive(1625) pre(1342) sample(1324) music(1302) demo(1144) download(988) audio(913) track(867) drum(848) guitar(834) file(830) clip(798) code(774) perform(765)

Table 2. Common Tokens from External Metadata

5.2. Similarity Measures

	TFIDF	Term Prob.	Doc. Overlap
Link Text	0.28	0.25	0.23
Surrounding Text	0.37	0.30	0.27
Combined	0.42	0.36	0.32

Table 3. General Similarity Comparisons

Table 3 shows similarity results for each of the three analyzes types, comparing sound file name tokens with other "visible" text sources.

Table 4 shows the same results between the entire document of combined information and each subtype, divided by sound file type.

6. CONCLUSIONS

Sound files on the web are sparser than other types of information, and occur less frequently than than the number of web pages parsed. However, sound files are not evenly distributed across all pages, but tend to exhibit a strong hub structure [12], with particular sources often leading to multiple files.

	TFIDF	Term Prob.	Doc. Overlap
WAVE			
filename	0.67	0.56	0.56
anchor	0.25	0.17	0.17
surrounding	0.45	0.33	0.32
AIFF			
filename	0.68	0.44	0.43
anchor	0.49	0.27	0.27
surrounding	0.57	0.52	0.51
MPEG-1			
filename	0.66	0.46	0.44
anchor	0.43	0.32	0.32
surrounding	0.52	0.46	0.45

Table 4. Similarity Comparisons by File Type

Pages with sound files tend to feature many of the same text tokens. For a web crawler, page retrievals based on these tokens is more likely to produce web pages that have links to sound files, and can constitute the basis for a priorization scheme for focused crawling of audio information. Also, there are differences in token patterns and distribution between different file types, and can be taken as an indicator of differences in usage patterns for different types.

There are correspondences between sound file names and associated visible text tokens on web pages containing links to sound files. The greatest amount of correspondence occurs when both the surrounding text and visible anchor text are retrieved and parsed, indicating that the use of both types of information should be used to produce better metadata descriptions of linked sound files than the use of the sound file names alone.

7. FUTURE WORK

This research is intended to be used in the improvement of the AROOOGA search engine and web crawler. Future crawls testing the value of this information for focused crawling are planned. The results detailed here also need to be tested against a larger portion of the web. One of the next steps in this line of research is to undertake a much larger crawl, possibly necessitating the use of a different class of machinery.

One of the functions of AROOOGA is to combine DSP and audio analysis techniques with the crawling process to improve the quality of information available to search engine users. Additional research will involve comparison of DSP analyzes of retrieved sound files with the textual information from the associated web pages, particularly in the study of genre and web localization.

8. REFERENCES

[1] E. Amitay. Using common hypertext links to identify the best phrasal description of target web documents. In *Proceedings of the SIGIR'98 Post-Conference Workshop on Hypertext Information Retrieval*. n.p., 1998.

[2] P. Boldi, B. Codenouti, M. Santini, and S. Vigna. Ubicrawler: A scalable fully distributed web crawler. In *Proceedings of the Eighth Australian World Wide Web Conference*, page n.p., 2002.

[3] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Proceedings of the 7th International World Wide Web Conference*, pages 107–17, 1998.

[4] R. Burke. Salticus: guided crawling for personal digital libraries. In *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pages 88–89, 2001.

[5] S. Chakrabarti, M. Berg, and B. Dom. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks (Amsterdam, Netherlands: 1999)*, 31(11–16):1623–1640, 1999.

[6] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource list compilation by analyzing hyperlink structure and associated text. In *Proceedings of the 7th International World Wide Web Conference*, pages 65–74, 1998.

[7] B. D. Davison. Topical locality in the web. In *Proceedings of the 23rd Annual Conference on Research and Development in Information Retrieval*, pages 272–79, 2000.

[8] J. Dean and M. Henzinger. Finding related pages in the World Wide Web. *Computer Networks*, 31(11–16):1467–79, 1999.

[9] J. Edwards, K. McCurley, and J. Tomlin. An adaptive model for optimizing performance of an incremental web crawler. In *Proceedings of the Tenth International World Wide Web Conference*, pages 106–113, 2001.

[10] V. Harmandas, M. Sanderson, and M. D. Dunlop. Image retrieval by hypertext links. In *Proceedings of the 20th International Conference on Research and Development in Information Retrieval*, pages 296–303, 1997.

[11] A. Heydon and M. Najork. Mercator: A scalable, extensible web crawler. *World Wide Web*, 2(4):219–29, 1999.

[12] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–32, 1999.

[13] M. Najork and J. L. Wiener. Breadth-first crawling yields high-quality pages. In *Proceedings of the 10th International World Wide Web Conference*, pages 114–8, 2001.

[14] F. Pachet and D. Laigre. A naturalist approach to music file name analysis. In *International Symposium on Music Information Retrieval*. 51–8, 2001.

[15] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[16] N. Rowe. Marie-4: a high-recall, self-improving web crawler that finds images using captions. *Intelligent Systems, IEEE*, 17(4):8–14, 2002.

[17] S. Sclaroff, L. Taycher, and M. La-Cascia. Imagerover: a content-based image browser for the world wide web. In *IEEE Workshop on Content-based Access of Image and Video Libraries*, pages 2–9, 1997.

[18] V. Shkapenyuk and T. Suel. Design and implementation of a high-performance distributed web crawler. In *Proceedings of the 18th International Conference on Data Engineering*, pages 249–54, 2002.

[19] H. Vinet, P. Herrera, and F. Pachet. The Cuidado project: New applications based on audio and music content description. In *Proceedings of the International Computer Music Conference*, pages 450–454. ICMA, 2002.

[20] H. Yan, J. Wang, X. Li, and L. Guo. Architectural design and evaluation of an efficient Web-crawling system. *Journal of Systems and Software*, 60(3):185–193, 2002.