

Geospatial Location of Music and Sound Files for Music Information Retrieval

Ian Knopke

McGill Music Technology
Montréal, Québec, Canada

ian.knopke@mail.mcgill.ca

ABSTRACT

A relatively new avenue of Web-based information retrieval research, intended to semantically improve information extraction, is the idea of using geographical information to accurately locate resources. This paper introduces a technique for locating sound and music files geographically. It uses information extracted from the Web relating to audio resources and combines it with geospatial location data to provide new information about audio usage in various countries. The results presented here illustrate the enormous potential for MIR to use the vast amount of audio materials on the Web within a physical and geographical context. Statistics of audio usage around the world are provided, as well as examples of other applications of these techniques.

Keywords: ISMIR, AROOOGA, geospatial, mapping, World Wide Web, Web crawler, GIS, semantic, CIA, McGill, music

1 INTRODUCTION

The World Wide Web is the largest data repository on Earth, constituting billions of pages of textual information, as well as providing access to many other types of resources, including multimedia. Many of the currently available search techniques are insufficient for answering specific queries about these resources. Searches for textual information may work well under many circumstances, but current techniques often fail to take into account the specific nature of other media, or do not meet the needs of more discerning users.

A relatively new avenue of Web-based information retrieval research, intended to semantically improve information extraction, is the idea of using geographical information to accurately locate resources. While the Web is usually considered to be a “virtual” space, without dis-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

©2005 Queen Mary, University of London

tance or dimension, it is rapidly becoming possible to determine the exact physical location of Internet resources in the real world. This has many potential possibilities for information retrieval. For instance, search engines could be improved by ranking local user resources for common queries (stores, movie theaters, etc.) ahead of more distant locations. Geospatial data mining techniques could also prove useful to market analysts in targeting new products for potential markets (Buyukokkten et al., 1999). Other uses, such as spatial Web browsing, verifying user locations, or even preventing credit card fraud have been proposed (MaxMind, 2005).

This paper introduces a technique for locating sound and music files geographically. It uses information extracted from the Web relating to audio resources and combines it with geospatial location data to provide statistics about audio usage in various countries around the world. The results presented here illustrate the enormous potential for MIR to make use of the vast amount of audio materials on the Web within a physical and geographical context.

The availability of such information can be exploited in a number of ways:

Geospatial Music Web Browsing The possibility of applying this information to mapping systems is especially promising, and could form the basis for a sound file browsing system, especially if used in conjunction with other MIR concerns such as genre information or melodic detection techniques, and combined with a Web-based dynamic mapping system such as that recently introduced by Google (2005).

Music Marketing Accurate information about the nature of sound and music in a particular geographical location could form an important resource for the commercial music sector as well, for planning music marketing strategies, or even in the selection of touring locations for musical ensembles.

Bandwidth Optimization User downloads of music require considerable Internet bandwidth resources. Geographical information could be used to optimize bandwidth and download times by pre-selecting sites that are geographically closer to users, especially in the case of similar or duplicated content. This could have important ramifications for Web-based digital

library systems or Internet sound repositories that exchange audio files.

“Musicological” Research Obtaining information about current music listening trends can be extremely difficult, and is often limited to statistics reported by industry sources. For some locations around the world, such information may not be available at all. The technique presented here provides an alternate method of independently gathering and verifying this type of information.

2 GEOSPATIAL MAPPING

Building a geospatial database of domains and IP addresses is accomplished through the use of multiple data sources that may in themselves be incomplete, but can be combined to form a more consistent whole. Sources can be categorized into two main types: available information about host machines, or from Web page content stored on specific servers.

Host machine information, when present, can be obtained directly from the `Whois` database for domain name registrations. Routing information can also be used, such as monitoring the last hop provided by the `traceroute` utility. Some information is available from Domain Name System (DNS) servers, although this considered to be less accurate geographically.

Other information can be gathered by examining Web page content on specific servers. Information such as addresses, postal codes, place names and even telephone numbers can provide valuable clues as to the real location of a host machine. McCurley (2001) provides a good overview of geospatial database-creation techniques.

Much geospatial Web research is undertaken in the context of semantic Web searching (Hiramatsu and Reitsma, 2004; Jones et al., 2002). Egenhofer (2002) notes that, for the current Web to become a true information resource, it requires better semantic search techniques and proposed a new research agenda for the creation of a *Semantic Geospatial Web* that coordinates existing search engine query information with geospatial data. (Tomko, 2002) has provided an assessment and user study of the suitability of the Web for providing information for common navigational tasks.

3 THE AROOOGA MIR SYSTEM

AROOOGA (Articulated Resource for Obsequious Opinionated Observations into Gathered Audio) is a Music Information Retrieval system designed to locate and analyze audio resources on the World Wide Web (Knopke, 2004a, 2005). Unlike other Internet-based sound distribution systems such as the Kazaa P2P program or Apple’s iTunes, AROOOGA only gathers information about sound and music files that are available on public Web pages. These pages must be easily accessible; pages behind sign-up forms, password protection, or pay-to-download schemes are not generally retrievable by automatic means. As these files are made freely available to anyone with an Internet connection, these resources are

generally not encumbered by the kind of commercial restrictions that may arise with other systems.

The central component of AROOOGA is a high-capacity, scalable, distributed Web crawling system. The secret to AROOOGA’s speed is a set of *retriever* modules distributed across many computers, making it possible to retrieve multiple resources in parallel. This is coordinated by a central managing component known as the *crawl manager* that controls the list of URLs to be retrieved, as well as handling all data storage. Inter-machine communication is handled by a customized message-passing protocol over Ethernet connections.

AROOOGA begins a crawl by working from a seed list, usually of less than a hundred URLs. After retrieval, each Web page encountered is mined for two types of information: links to other Web pages or audio files, and Web page text relating to linked sound files. Links to other pages are then added to a queue for later retrievals, allowing the crawling process to continue. Linked sound files are also retrieved and analyzed. Information extracted for each audio file falls into one of three categories: external metadata gathered from the linking Web page, internal metadata stored within the sound file such as sampling rate or keywords, and information gathered from analysis of the actual encoded audio. In effect, the external metadata acts as a kind of annotation system for the retrieved sound files (Knopke, 2004b), and the association of the three types of information has been shown to provide better indexing than can be achieved from mining Web pages or audio files alone (Knopke, 2005). For audio extraction, AROOOGA currently uses the MARSYAS (Tzanetakis and Cook, 2000) system, primarily for music/speech determination and to obtain the genre of music files (Tzanetakis et al., 2000).

Unlike other general-purpose Web crawling or search engine systems such as Google or Yahoo, information extraction is done at the time of analysis and the results are sent to the crawl manager for permanent storage, and later use in resolving queries. Only the results of the three types of analysis are stored, and all other textual information on retrieved Web pages is considered irrelevant and ignored. This allows the system to focus solely on music and sound resources in order to answer specific audio-related queries, forming the core of a search engine. The system is programmed in Perl, with audio analysis components in C and C++ for greater efficiency.

Audio data on the Web is quite sparse as compared to the number of Web pages. The majority of sound files on the Web tend to be MP3 types, with smaller numbers of WAVE and AIFF files occurring regularly. Using statistical data taken from several sample crawls, it is possible to extrapolate some information about the number and nature of sound resources on the Web. While it is difficult to estimate the exact size of the Web, by using a plausible size of 9 billion pages we can arrive at the following values, shown in Table 1.

It is clear that there is an exceptional amount of sound and music materials available on the Web, although much of this is difficult to search for with any accuracy.

Table 1: Predicted Values for a Web of 9 Billion Pages

Type	% of Total	Number
Web pages	100.00	9,000,000,000
Pages with audio links	0.26	23,400,000
Sound files	2.39	215,100,000
WAVE	31.99	68,810,490
AIFF	1.63	3,506,000
MPEG-1	66.33	142,675,830
OTHER	0.06	129,060

4 METHODOLOGY

A sample crawl was undertaken of approximately 600,000 Web pages. The seed list was chosen to reflect typical pages a user might encounter within a more narrow range of audio-related interests, and can be seen as a form of seed pre-selection or focused crawling (Chakrabarti et al., 1999; Ester et al., 2001). Starting URLs of Web sites were used that specifically contained multiple links to sound files, many of them .com domains of music companies or portals. Duplicate sound files were removed during the crawling process. Some examples of sites used are shown in Table 2.

Table 2: Partial URL seed list

www.loopmasters.com
www.cobwebaudio.co.uk
www.pro-music.org
music.download.com
www.soundcentral.com
www.analoguesamples.com
www.multi-edit.com
www.musicstuff.de
free-music.com/fma2000
www.propellerheads.se

The country of origin where the sound file resides was determined using a geospatial database from MaxMind corporation, claimed to have a 97% accuracy rate for country determination (MaxMind, 2005). These were also correlated against additional information taken from the CIA World Factbook (CIA, 2004). Map plotting was done using the Generic Mapping Toolkit (Wessel and Smith, 2005). All results were calculated using custom Perl code, and geospatial database access was provided through the Geo::IP Perl Module (Mather, 2005).

5 RESULTS

General statistics for the crawl are shown in Table 3. Attempted downloads refers to the total number of requests for Web pages. Some pages in any crawl will be unretrievable, usually due to broken links or security features, and are listed as unsuccessful downloads. By dividing the number of retrieved sound files by the number of successful page downloads, an average of the number of sound file resources per page can be calculated (0.07).























The actual country of origin for each audio file was derived by querying each URL against the geospatial

Table 3: General crawl statistics

Page Links collected	85,764,415
Attempted downloads	612,311
Successful downloads	532,849
Unsuccessful downloads	79,462
Success/Unsucc. %	87.02%
Unique sound files	37,353
Sound / Page	0.07

database. Table 4 give the top geospatial statistics listed by country. Approximately three quarters of all the sound files retrieved (71.86%) are located in the United States, regardless of domain suffix. Italy (10.35%), Germany (4.11%) and Great Britain (3.72%) were also found to contain significant sound file repositories.

Table 4: Top statistics by country

Country	Count	%
USA 	26842	71.86%
Italy 	3866	10.35%
Germany 	1537	4.11%
Great Britain 	1390	3.72%
Japan 	378	1.01%
Canada 	335	0.90%
South Korea 	253	0.68%
France 	241	0.65%
Spain 	199	0.53%
Sweden 	159	0.43%
Belgium 	100	0.27%
Australia 	94	0.25%
Austria 	92	0.25%
Netherlands 	85	0.23%
Denmark 	68	0.18%
Switzerland 	52	0.14%
Czech Republic 	48	0.13%
Russia 	35	0.09%
South Africa 	34	0.09%
Finland 	31	0.08%
Poland 	26	0.07%
Norway 	19	0.05%
Other	37	0.10%
Unknown	1432	3.83%
Total	37353	100%

This information can also be grouped together into larger geographical regions. Table 5 gives these results for each continent, correlated by the three main file types encountered: WAVE, AIFF, and MP3. By far the most common type of sound files are MP3 files, at 81.58%, with the exception of the Asian continent, where WAVE files figure more predominantly. This may indicate a possible preference for operating system and hardware in that region.

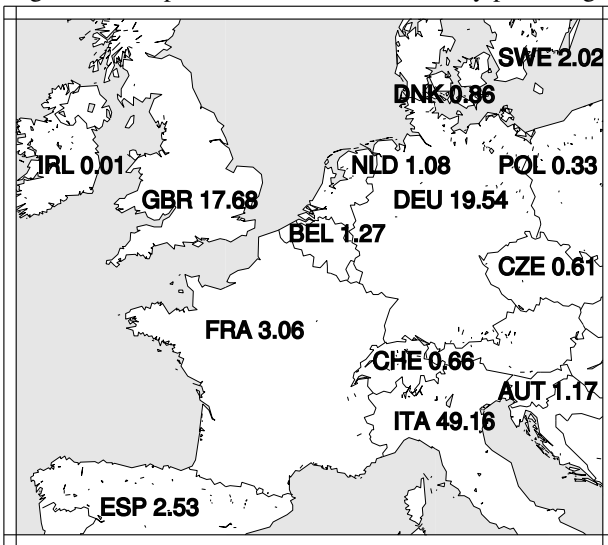
Access to geospatial information naturally suggests the possibility of displaying such statistics using a geographical mapping system. Figure 1 shows an example of

Table 5: Sound file types sorted by continent

Continent	WAVE %	AIFF %	MP3 %
Africa	0.00	0.00	100.00
Asia	89.22	0.16	10.62
Australia	1.06	1.06	97.87
Europe	4.83	0.92	94.26
N America	17.03	3.50	79.45
S America	0.00	0.00	100.00
All	15.55	2.86	81.58

such a distribution for the European region, and demonstrates how such information might be used in combination with a Web-based graphical interface for file retrieval.

Figure 1: European sound file distribution by percentage



Knowing the geographic location of a file makes it possible to correlate this information with other statistics, such as the size of the population or the area of the country, and could be used as a kind of data normalization. For instance, a country with a smaller population would be expected to have fewer sound and Internet resources available. Table 6 and 7 gives such figures for the population and area respectively.

6 FUTURE WORK

These results are considered somewhat preliminary, and need to be applied to larger crawls of the Web. Currently this is limited by the available search hardware. Also, the geospatial database used for this study can only identify the country within which the host server resides. More accurate databases, with resolution to the level of cities or even street address, are available on a subscription basis. The use of improved resources would allow for additional possibilities, such as accurate geographical distance measurements between server locations or could be used in conjunction with other semantic search or social networking techniques (Hiramatsu and Reitsma, 2004).

Table 6: Density of sound files per person

Country	Files	Pop.(10 ⁶)	Density(10 ⁻⁶)
USA	26842	278.06	96.53
Italy	3866	57.68	67.03
Great Britain	1390	59.65	23.30
Germany	1537	83.03	18.51
Sweden	159	8.88	17.92
Denmark	68	5.35	12.70
Austria	92	8.15	11.29
Canada	335	31.59	10.60
Belgium	100	10.26	9.75
Switzerland	52	7.28	7.14

Table 7: Density of sound files per square Km

Country	Files	Area(10 ⁶)	Density(10 ⁻⁶)
Italy	3866	0.30	12834.05
Great Britain	1390	0.24	5677.64
Germany	1537	0.36	4305.07
Belgium	100	0.03	3277.61
USA	26842	9.63	2787.59
South Korea	253	0.10	2569.05
Netherlands	85	0.04	2046.61
Denmark	68	0.04	1577.95
Japan	378	0.38	1000.44
France	241	0.55	440.56

This information is also intended to form the basis for a spatially-aware music browsing system of the Web, making it possible to “point-and-click” on specific geographical maps and obtain a list of resources by country or city. Even finer resolutions should be possible.

7 CONCLUSIONS

The use of geospatial information in combination with sound and music information on the World Wide Web shows much promise, and has the potential to prove useful for many different information retrieval situations. With a sufficiently large pool of data, the ability to geographically locate audio resources can be used to derive information that is typically difficult to obtain by other means. This data can also be used in conjunction with other statistical information to provide new measurements of sound and music and could form an important new resource for semantic Web search techniques and MIR research in general.

8 REFERENCES

References

- O. Buyukokkten, Ji. Cho, and N. Shivakumar H. Garcia-Molina, L. Gravano. Exploiting geographical location information of web pages. In *Proceedings of Workshop on Web Databases*, pages 91–6. ACM Press, 1999.
- S. Chakrabarti, M. Berg, and B. Dom. Focused crawling:

- A new approach to topic-specific web resource discovery. *Computer Networks*, 31(11-6):1623-40, 1999.
- CIA. World factbook, 2004. <http://www.cia.gov/cia/download.html>.
- M. Egenhofer. Toward the semantic geospatial web. In *ACM-GIS. GIS*, 2002.
- M. Ester, M. Gross, and H. P. Kriegel. Focused web crawling: A generic framework for specifying the user interest and for adaptive crawling strategies. In *Twenty-Seventh International Conference on Very Large Databases*, 2001.
- Google. Maps, 2005. <http://maps.google.com/>.
- K. Hiramatsu and F. Reitsma. Georeferencing the semantic web: ontology based markup of geographically referenced information. In *Joint EuroSDR/EuroGeographics workshop on Ontologies and Schema Translation Services*, 2004.
- C. Jones, R. Purves, A. Ruas, M. Sanderson, M. Sester, M. van Kreveld, and R. Weibel. Spatial information retrieval and geographical ontologies: An overview of the spirit project. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 387-8, 2002.
- I. Knopke. AROOOGA: An audio search engine for the World Wide Web. In *Proceedings of the International Computer Music Conference*, pages 290-3, November 2004a.
- I. Knopke. Sound, music and textual associations for the World Wide Web. In *Proceedings of the International Symposium on Music Information Retrieval*, pages 484-8, October 2004b.
- I. Knopke. *Building a Search Engine for Music and Audio on the World Wide Web*. PhD thesis, McGill University, 2005.
- T. Mather. Geo::ip perl module, 2005. <http://search.cpan.org/~tjmather/Geo-IP-1.25/>.
- MaxMind. Geoip free country database, 2005. <http://www.maxmind.com/download/geoip/database/>.
- K. McCurley. Geospatial mapping and navigation of the web. In *WWW '01: Proceedings of the tenth international conference on World Wide Web*, pages 221-9. ACM Press, 2001.
- M. Tomko. Case study - assessing spatial distribution of web resources for navigation services. In *4th International Workshop on Web and Wireless Geographical Information Systems*, 2002.
- G. Tzanetakis and P. Cook. MARSYAS: A framework for audio analysis. *Organized Sound*, 4(3):169-75, 2000.
- G. Tzanetakis, G. Essl, and P. Cook. Automatic musical genre classification of audio signals. In *International Symposium on Music Information Retrieval*. n.p, 2000.
- P. Wessel and W. Smith. Generic mapping tools, 2005. <http://gmt.soest.hawaii.edu/>.