Informetrics and Music Information Retrieval: An Informetric Examination of a Folksong Database

J. Stephen Downie
Faculty of Information and Media Studies
University of Western Ontario
jdownie@julian.uwo.ca

This study examined the informetric properties of intervallic n-grams as derived from a database of 9354 folksongs. The informetric properties of the melodies were compared and contrasted to those of traditional text. Understanding these similarities/dissimilarities can play a vital role in the creation of a successful Music Information Retrieval (MIR) system. Various styles of n-gram creation were examined some of which exhibited great promises. Finally, this paper shows how the informetric analyses can be used to give theoretical justifications for the application of traditional IR methodologies in constructing MIR systems.

Background

There seems to be as many approaches to developing Music Information Retrieval (MIR) systems as there are developers. Some have designed complex suites of computer tools to analyse all the varied facets of music (e.g., Huron 1991). Others have tried to automate the thematic catalogue by including *incipits* (i.e., beginning phrases) as part of a bibliographic record (e.g., RISM 1997). Still others have explored the idea of using sophisticated approximate string-matching techniques (e.g., Ghias, Chamberlin and Smith 1995; McNab et al. 1996). One thing that unites all of these approaches is that they have some kind of shortcoming. The more powerful analytic systems can be very difficult to use, *incipit* indexes leave out large amounts of music that might be of interest, and approximate string-matching can be computationally expensive without necessarily giving better results (Downie 1997).

Taking our cue from those thematic catalogues that have reduced the amount of musical information represented (e.g., Keller and Rabson 1980) we have been developing a prototype MIR system (Tague, Downie and Dunne 1993; Downie 1994) based upon the intervals (i.e., the signed differences between pitches) found within monophonic melodies (Downie 1995, 1997a). We believe that there is enough information contained within an interval-only representation of monophonic melodies that effective retrieval of musical information can be achieved. We aim to extend the thematic catalogue model by affording access to melodic segments found anywhere within a melody. To achieve this extension we fragment the melodies into length-n subsections called n-grams. The length of these n-grams and the degree to which we precisely represent the intervals are variables to be analysed.

For example, consider Figure 1 where Ix represents an interval "I" at location *x* in a melodic string. The **Contiguous String** represents the sequence of intervals as extracted from a melody. The next three items show how we partition the **Contiguous String** into substrings of length-4, 5, and 6, called 4-grams, 5-grams, and 6-grams, respectively. In our nomenclature we denote these n-grams as L4, L5 and L6, respectively.

Figure 1. N-grams of Intervallic Information.

Contiguous string:

I I I I I I I I I I ...
1 2 3 4 5 6 7 8 9 10 11

4-gram representation:

[I I I I] [I I I I] [I I I I] [I I I I] [I I I I] [I I I I] [I I I I]...
1 2 3 4 2 3 4 5 3 4 5 6 4 5 6 7 5 6 7 8 6 7 8 9 8 9 10 11

5-gram representation:

[I I I I I] [I I I I I] [I I I I I] [I I I I I] [I I I I I] [I I I I I] [I I I I I]...
1 2 3 4 5 2 3 4 5 6 3 4 5 6 7 4 5 6 7 8 5 6 7 8 9 6 7 8 9 10 7 8 9 10 11

6-gram representation:

[I I I I I I] [I I I I I I] [I I I I I I] [I I I I I I] [I I I I I I] [I I I I I I]...
1 2 3 4 5 6 2 3 4 5 6 7 3 4 5 6 7 8 4 5 6 7 8 9 5 6 7 8 9 10 6 7 8 9 10 11

Depending on how precisely one wished to represent the intervals in a melody, "I" can take on many different value ranges. If one wanted to represent only the direction of the intervals one could group all the intervals into one of three classes (e.g., Same, Up, and Down). In our nomenclature we call such a classification scheme

C3, where "3" indicates the number of elements in the scheme. The C3 scheme is considered very forgiving because a user would only have to remember the contour of a melody and not its exact intervals. The most precise classification scheme is CU (Unclassified) where each of the intervals is represented as found in the melody. The CU scheme is not very forgiving but it does give an accurate representation of the melodic lines. On a recall-precision continuum, C3 is designed to enhance recall (at the expense of precision) while CU is designed to enhance precision (at the expense of recall).

These n-grams form discrete units of melodic information in much the same manner as words are discrete units of language. Thus, we have come to consider them "musical words." We hypothesize that, for the purposes of music information retrieval, we can treat them as "real words" and thus apply traditional text-based information retrieval techniques. We further hypothesize that there might be some type of equivalency between the intervals found within the n-grams and letters found within "real words."

In Downie (1997) we examined the informetric characteristics of four n-gram length/classification scheme (C*x*L*x*) combinations as applied to a toy collection of 100 folksongs: C7L6, C12L5, C14L5 and C23L4. The study suggested that "musical words" are similar to "real words" in that their distributions of occurrence can both be modelled using the Zipf or Lotka functions. That "musical words" seem to carry more information than "real words" was also suggested. The small size of the database investigated, however, precluded generalization.

Introduction

When performing informetric analyses it is customary to make the distinction between *types* and *tokens*. Types are the unique entities (*e.g.*, words, letters, n-grams, intervals, *etc.*) that make up a corpus of interest. Tokens are the instances of the types. For example, the collection [dog, dog, cat] contains two types (dog and cat) and three tokens. In this study the types are the distinct n-grams (or intervals) and the tokens the instances of same.

To determine in what ways intervals and letters, n-grams and words, might —or might not— be similar, and to predict retrieval performance, descriptive data was collected in an effort to answer the following questions:

1. How many types* (and tokens) are present in the database (* i.e., size of the "alphabet" in the case of interval; size of "vocabulary" in the case of the n-grams)?
2. How are the tokens distributed over the types?
3. How are the types (and tokens) distributed over the songs?
4. How much information is contained, on average, in the types.
5. What effect does the classification (C) and n-gramming (L) processes have on the answers to the questions above.

The information presented in this paper represents a portion of the informetric analyses section of the author's proposed doctoral thesis, *Evaluating a simple approach to music information retrieval: Representing melodies as collections of "musical words"* (Downie 1997). As a doctoral thesis the parameters under investigation are substantially more ambitious than those referred to in the author's prior work[1]. For example, rather than a toy database of 100 songs, this study used the database collection of 9354 folksongs created by McNab (McNab et al. 1996). The McNab collection contains a mix of folksongs that represent American, German, Irish and Chinese traditions. With this much larger database, and the wide range of musical traditions it comprises, it is hoped that conclusions drawn will be more generalizable (at least within the domain of folksongs and simpler vocal music).

The choice of n-gram lengths (L4, L5, and L6) examined has not changed from the Downie (1997a) study. However, the characteristics of the classification schemes are different from those evaluated in earlier papers (Downie 1995; 1997a). The C7 and C12 classification schemes used in the previous studies did not include directional

information (i.e., whether the interval went up or down). Dowling (1978) has shown that one of the strongest factors in melodic memory and recognition is the shape, or contour, of a tune. That is to say that users might not remember the specific pitches of a given song but they do have a strong sense of the direction of its intervals. For this reason, all classification schemes in the present study retain directional information.

Furthermore, for reasons of statistical significance Kinnucan (1996) has rightly suggested that the number of elements in each classification scheme be as distinct as possible. An important factor under investigation in this study is the effect of the size of the classifactory set Cx on retrieval effectiveness, where x is the number of elements in the set. The present C3, C7, C15, and CU set sizes were chosen because they are substantially different from each other (unlike C&, C12, C14, and C23 of the earlier studies. Table 1 presents the twelve combinations of the C*x*L*x* factors examined.

Table 1. N-gram databases

C3L4	C7L4	C15L4	CUL4
C3L5	C7L5	C15L5	CUL5
C3L6	C7L6	C15L6	CUL6

With any classification scheme the amount of information lost through the collapsing of data into categories is inversely proportional to the number of categories available. It is possible to envision the C3, C7, C15, and *CU* sets as lying evenly spaced upon an information-loss continuum with C3 representing "high" information loss, C7 "medium" information loss, C15 "low" information loss, and C23 "no" information loss. Thus, an examination of the informetric characteristics of the Classification C factor becomes an examination of the effect of information loss on potential retrieval effectiveness.

A precision-recall continuum also exists for the n-gram length factor (L*x*). Shorter n-gram lengths produce fewer unique strings (lower precision) while longer n-grams produce many more unique strings (higher precision). The number of potential unique strings (i.e., "vo-

cabulary" size) varies exponentially with the length of the n-gram according to the equation:

$$\text{Potential number of unique strings} = C^L,$$

where C is the size of the classificatory set; and, L the n-gram length.

Method

A series of PERL programmes written by the author and his assistant, Kevin Kennedy, were used to manipulate and analyse the music information contained within the McNab database. Because the McNab database comprises melodic representations that include both pitch and rhythm information, reducing the folksongs to their interval-only representation was the first task. A baseline database (BD) file, where each song was represented by a contiguous string of unclassified intervals, was thus created. The occurrence of each interval type within the BD file was counted. The relative frequency of each interval type was also determined (Table 2, middle columns).

The CU classification scheme was created using the intervals as they occurred in the BD file. However, to make the resulting n-grams easier to read, positive intervals were converted to their corresponding upper-case alphabetic characters. Negative intervals were converted to lower-case alphabetic characters. The interval *0* (i.e., note repeats) was assigned the character *a*.

The remaining classification schemes (Table 2) were created with two constraints in mind:

1. the directional information had to be preserved; and,

2. the resulting groupings had to maximize average information (i.e., entropy) according to Shannon's (Shannon and Weaver 1949) equation:

$$\overline{H} = -\sum_{r=1}^{n} P_r \log_2 P_r$$

where $\overline{H}$ is the entropy of a collection of n types r,

and p_r is the probability of occurrence of type r.
The unit of measurement for entropy is the *bit*.

Entropy was maximized, within the given prior constraint of directionality, in an effort to minimize the potentially detrimental effects of information loss caused by the classification process. After the classification schemes had been determined, the n-grammed databases found in Table 1 were created.

Data was next collected about the occurrence of intervals across the songs using the BD file (Table 3). Interval entropy for the BD file was calculated as 3.3891 bits. Interval entropy was then calculated for the intervals represented within the n-gram databases and compared with various theoretic, baseline, and published values (for "real letters") , the significance of which will be discussed in due course (Table 4).

Attention was then turned to examining the informetric properties of the n-grams. Data about the number of tokens in each database and their frequency of occurrence in each song was collected and is summarized in Table 5. The number of n-gram types present in each database was counted and the potential number of n-gram types calculated for each C*x*L*x* combination (Table 6). Data about the distribution of types across the databases and across the songs[2] was collected. Table 7 presents this information together with the calculation of n-gram entropies. A byproduct of the analysis that examined the frequency of n-gram type occurrences across the songs was the calculation of the probabilities that a given type, chosen at random from the list of types found within a database, would occur in x or fewer songs, with x ranging from 1 through 20 (Table 8). Worst-case scenarios for each database , wherein the individual n-gram types that occurred in the most songs were determined, were the last set of analyses run (Table 9).

Table 2. Summary of Interval Occurrence and Classification Schemes

Negative Interval Classification Codes					Positive Interval Classification Codes				
C3	C7	C15	CU	-Interval (Probability)	+Interval (Probability)	CU	C15	C7	C3
a	a	a	a	0 (0.2056)	0 (0.2056)	a	a	a	a
b	c	b	b	-1 (0.0603)	1 (0.0473)	B	B	C	B
b	b	c	c	-2 (0.2093)	2 (0.1491)	C	C	B	B
b	c	d	d	-3 (0.0792)	3 (0.0640)	D	D	C	B
b	d	e	e	-4 (0.0264)	4 (0.0265)	E	E	D	B
b	d	f	f	-5 (0.0297)	5 (0.0480)	F	F	D	B
b	d	g	g	-6 (0.0010)	6 (0.0004)	G	G	D	B
b	d	g	h	-7 (0.0125)	7 (0.0132)	H	G	D	B
b	d	g	i	-8 (0.0043)	8 (0.0034)	I	G	D	B
b	d	g	j	-9 (0.0028)	9 (0.0061)	J	G	D	B
b	d	g	k	-10 (0.0015)	10 (0.0030)	K	G	D	B
b	d	g	l	-11 (0.0000)	11 (0.0001)	L	G	D	B
b	d	g	m	-12 (0.0017)	12 (0.0037)	M	G	D	B
b	d	g	n	-13 (1.19E-05)	13 (0.0000)	N	G	D	B
b	d	g	o	-14 (0.0001)	14 (0.0002)	O	G	D	B
b	d	g	p	-15 (3.58E-05)	15 (0.0001)	P	G	D	B
b	d	g	q	-16 (2.39E-05)	16 (0.0001)	Q	G	D	B
b	d	g	r	-17 (2.58E-05)	17 (0.0001)	R	G	D	B
b	d	g	t	-19 (3.98E-06)	19 (3.18E-05)	T	G	D	B
b	d	g	v	-20 (7.95E-06)	20 (5.96E-06)	U	G	D	B
					21 (3.98E-06)	V	G	D	B
					22 (3.98E-06)	W	G	D	B
					24 (3.98E-06)	Y	G	D	B

Table 3. Descriptive Data about Intervals

	Interval Data (per song)	Unique Intervals (per song)
Mean	53.78	10.5
SD	29.37	2.27
Range	494	21
Minimum	7	3
Maximum	501	24
Total	503,086	
Entropy	3.389 bits	

Table 4. Interval Entropy Values Derived from n-grammed Databases

	F0	F1	F1-F0	F1(BD)-F1(n-grammed)	F1(Shannon)-F1(n-grammed)
C3L4	1.585	1.5224	-0.0625	-1.8666	2.6176
C3L5	1.585	1.5225	-0.0624	-1.8666	2.6175
C3L6	1.585	1.5228	-0.0621	-1.8662	2.6172
C7L4	2.8074	2.7343	-0.0731	-0.6548	1.4057
C7L5	2.8074	2.7346	-0.0728	-0.6545	1.4054
C7L6	2.8074	2.7347	-0.0726	-0.6543	1.4053
C15L4	3.9069	3.279	-0.6279	-0.1101	0.861
C15L5	3.9069	3.2795	-0.6274	-0.1096	0.8605
C15L6	3.9069	3.2798	-0.6271	-0.1093	0.8602
CUL4	5.3923	3.3928	-1.9995	0.0037	0.7472
CUL5	5.3923	3.3939	-1.9984	0.0048	0.7461
CUL6	5.3923	3.3948	-1.9975	0.0057	0.7452

Table 5. Descriptive Data about n-gram Tokens

	L4	L5	L6
Num. Tokens	475,024	465,670	456,316
Average Tokens/Song	50.78	49.78	48.78
SD Tokens/Song	29.37	29.37	29.37
Maximum Tokens/Song	498	497	496
Minimum Tokens/Song	4	3	2

Table 6. Number of n-gram Types

	C3	C7	C15	CU
L4	81 (81)	2,298 (2401)	13,273 (20736)	21,796 (3,111,696)
L5	243 (243)	12,622 (16807)	50,954 (248832)	64,902 (1.31E+08)
L6	729 (729)	50,730 (117649)	126,346 (2985984)	139,428 (5.49E+09)
	Actual (Theoretic Max.)			

Table 7. Descriptive Data about n-gram Types

	C3L4	C3L5	C3L6	C7L4	C7L5	C7L6	C15L4	C15L5	C15L6	CUL4	CUL5	CUL6
Average Types/Song	26.19	32.75	36.62	38.04	40.12	40.97	39.06	40.68	41.33	39.2	40.77	41.4
SD Types/Song	8.65	13.15	16.83	16.9	19.67	21.3	17.87	20.22	21.67	18.03	20.32	21.75
Maximum Types/Song	70	129	190	192	250	285	209	260	300	212	260	303
Minimum Types/Song	4	3	2	4	3	2	4	3	2	4	3	2
Entropy(A) Types/Collection	6.02	7.5	8.97	9.98	12.25	14.38	11.53	13.88	15.8	11.78	14.08	15.92
Entropy (B) Types/Songs	6.18	7.63	9.07	10.06	12.31	14.43	11.66	13.97	15.88	11.94	14.19	16.01
Entropy (C) Theoretic Max	6.34	7.92	9.51	11.17	13.62	15.63	13.7	15.64	16.95	14.41	15.99	17.09
Entropy (B-A)	0.16	0.13	0.1	0.08	0.06	0.05	0.13	0.09	0.08	0.16	0.11	0.09
Entropy (C-A)	0.32	0.42	0.54	1.19	1.37	1.25	2.17	1.76	1.15	2.63	1.91	1.17
Entropy (C-B)	0.16	0.29	0.44	1.11	1.31	1.2	2.04	1.67	1.07	2.47	1.8	1.08
Entropy (A - Shannon)	-5.8	-4.32	-2.85	-1.84	0.43	2.56	-0.29	2.06	3.98	-0.04	2.26	4.1
Entropy (B-Shannon)	-5.64	-4.19	-2.75	-1.76	0.49	2.61	-0.16	2.15	4.06	0.12	2.37	4.19

Table 8. Probability that a given n-gram occurs in *x* or Fewer Songs

X	C7L4	C7L5	C7L6	C15L4	C15L5	C15L6	CUL4	CUL5	CUL6
1	0.03	0.13	0.29	0.26	0.40	0.57	0.40	0.50	0.62
2	0.06	0.21	0.44	0.37	0.56	0.73	0.53	0.65	0.77
3	0.09	0.27	0.55	0.45	0.65	0.81	0.61	0.73	0.84
4	0.11	0.32	0.62	0.50	0.70	0.86	0.66	0.78	0.88
5	0.13	0.36	0.68	0.54	0.75	0.89	0.70	0.81	0.91
6	0.14	0.39	0.72	0.57	0.78	0.91	0.72	0.84	0.92
7	0.16	0.42	0.75	0.60	0.80	0.92	0.75	0.85	0.94
8	0.17	0.45	0.78	0.62	0.82	0.94	0.76	0.87	0.95
9	0.19	0.47	0.80	0.64	0.84	0.94	0.78	0.88	0.95
10	0.20	0.50	0.82	0.66	0.85	0.95	0.79	0.89	0.96
15	0.25	0.58	0.88	0.72	0.90	0.97	0.84	0.93	0.98
20	0.28	0.65	0.92	0.76	0.92	0.98	0.86	0.95	0.98

Table 9. Number of Songs in which the Most Frequently occurring n-gram Type was Found

	C3	C7	C15	CU
L4	7349 (79)	2294 (25)	1891 (20)	1891 (20)
L5	4742 (51)	982 (10)	719 (08)	719 (08)
L6	2603 (28)	403 (04)	311 (03)	311 (03)
	Number of Songs (Percentage of Database)			

Observations and Discussion

Interval analyses

Average entropy is maximized when the probabilities of occurrence of each type are equal. In such a case, average entropy of a system can be expressed as:

$$F_0 = -\log_2 (1/n)$$

where F_0 is the average entropy of a collection of equally distributed n types. F_0 is also used to calculate the average entropy of a system where the number of types is known but the distribution of tokens over types is not. When the probabilities of occurrence *are known* for the types in a collection the OVERLINE H equation is used. The symbol F_1 is sometimes used to signify that observed probabilities were used in the calculations. Assuming an english alphabet of 26 characters (case excluded) the F_0 value of an english letter is 4.70 bits. Shannon, using observed letter occurrence probabilities, calculated the entropy of a letter as F_1= 4.14 bits (Shannon 1951).[3]

The drop in entropy estimates from 4.70 to 4.14 bits is caused by the fact that letter occurrence is not evenly distributed (i.e., the distributions are skewed).

The F_0 for the 42 intervals found in the baseline database was calculated as 5.39 bits. This value was greater than Shannon's because the number of interval types is greater than the number alphabetic characters. However, F_1 was calculated for the baseline database as 3.39 bits. Thus, intervals in the baseline database contain less information on average than letters do in "real" text. The discrepancy between baseline F_0 and F_1 values (2 bits) and between the baseline F_1 and the Shannon F_1 values (0.75 bits) can be attributed to the pronounced skew in the distribution of interval tokens over types in the baseline file. Note in Table 2 that the most frequently occurring type, (-2, or "b") represents nearly 21% of all the tokens in the database. Dewey (1950) calculated that the most frequently occurring letter type, "E", represented only 12.7% of letter tokens in english text. This increased skewing of the interval distributions is the reason why intervals contain less information than letters.

The mean number of intervals per song (53.78) along with the mean number of types per song (10.50) highlight some more differences between text and music (Table 3). First, that a complete "idea" could represented in such a compact manner is worth noting. One song was only 7 intervals long! Repetition of the complete melody as one cycles through a series of verses could exaggerate the significance of this compactness, however. Second, that on average only

10.50 interval types were used per song suggests a strong dissimilarity between intervals and letters. One would be hard-pressed to imagine a collection of "real" sentences that contained only 11 different letters on average.

The process of n-gramming under-represents those intervals that occur at the beginnings and endings of songs. The entropy of the intervals was calculated from the intervals counted from the n-grams to ascertain the amount of distortion caused by n-gramming. In Table 4, by reading down the column labelled $F_{1(BD)}$-$F_{1(n\text{-grammed})}$ and noting the differences in values between rows of similarity classified intervals, one can determine the amount of interval information lost (or gained) by n-gramming. The amount of distortion caused by n-gramming was extraordinarily slight (i.e. changing at the fourth decimal place).

The strongest cause of information loss was the application of the classification schemes,with C3 having an F_0 of only 1.58 bits while the F_0 for CU was 5.39 bits. The slight difference in C3's F_0 and F_1 values indicates that distribution of negative, neutral and positive intervals was relatively equal (with a slight skewing towards the negative intervals). Differences increased as C increased because the classification schemes were created with an eye to maximizing average interval entropy (i.e., shortening the tails of the distributions). Thus, CU with its "untrimmed" tail had the largest entropy drop. Finally, the differences found between the $F1_{(Shannon)}$ and the $F1_{(n\text{-grammed})}$ values indicate that C3 might not be suitable, and C7questionable, candidates for use in a music database that uses standard text retrieval methods. The sizes of their entropy discrepancies might be too large to overcome.

N-gram Analysis

Taking the mean number of tokens per song as 50 (Table 5) suggests that their might be some kind of equivalency between a folksong and a long title (or short abstract) in a traditional text database. Because the number of tokens is simply a function of the choice of n-gram length, little more need be said.

The number of n-gram types found in each database has important implications. The number of types counted in C3 databases is remarkably low, especially given the size of the databases. More significant is that the C3 databases have "saturated," or used up, all the available n-gram types (C7L4 and L5 are perilously close). In an inverted-file retrieval system, being "saturated" causes the length of postings lists to grow but not the length of the dictionary file. Such growth quickly undermines retrieval performance from both the efficiency (i.e., speed) and effectiveness (i.e., precision) standpoints. Within the context of the present database, having only 81"terms", or n-gram types, (C3L4) with which to retrieve a given song all but assures that retrieval performance will be unsatisfactory. Based solely on the data presented in Table 6, C7L5 appears to be the minimally acceptable C*x*L*x* combination.

The most interesting data in Table 7 is found in the entropy rows. Shannon's calculation of average word entropy was 11.82 bits (Heaps 1978). Taking this value as a benchmark indicates that the n-grams type created in the C7L5 and greater databases contained on average more information than that found in text (with CUL6 (15.92 bits) had the greatest advantage).

As suggested by the analysis of n-gram type counts, the C3L4 through C7L4 databases appear to be the weakest candidates for implementation given the significant negative differences between Shannon's values and those calculated from their respective data. The slight differences found in the **Entropy (B-A)** column indicates relatively insignificant influence that within-song redundancy had on determining the average information value of an n-gram.

The most interesting data of all concerned that which attempted to indicate what kind of retrieval performance to expect. Table 8 contains the data calculated to model the "best case" performances of the databases. "Best case" was defined as the probability that a randomly selected n-gram would occur in 20 or fewer songs, given that it is to be found in the list of that databases types. The number 20 was selected as a reasonable set size for browsing. C3 is not to be found in Table 8 as the "best" n-gram type (from C3L6) would re-

turn 83 songs if selected as a query. Not surprisingly, CUL6 was found to be the best with a 0.62 probability of returning only 1 song. A whopping 98% of the n-grams in CUL6 would return 20 or fewer songs! C15 databases were also strong. Again, however, the C7 databases, especially C7L4 and L5, appeared to be in a grey area with modestly acceptable results. From a worst-case standpoint, the CU and C15 databases also did best with identical results. C3L4 was the worst of the worst with one n-gram returning 79% of the database. C7 did not lag very far behind the leaders.

Summary and Conclusions

The hypothesized equivalencies between intervals and letters, and n-grams and words might be tenuous. However, as departure points for informetric analyses they have proven very useful. Based upon the calculation of average entropy values, a stronger argument can be made for the n-gram/word equivalency than for the interval/letter case. It was those databases whose n-grams exhibited equal, or greater, average information than text that showed the greatest potential for successful implementation in a MIR system modelled on standard text retrieval methods. Classification scheme was shown to be more influential than n-gram length. The C3 databases had very weak informetric properties. The results for the C7 scheme were inconclusive and suggests empirical retrieval evaluation before final judgement can be made. While the intended distinctions between the CU and C15 classification schemes were not as prominent as expected, both evaluated as the most promising of the classification schemes.

Acknowledgements

The author wishes to thank his supervisor, Dr. M. Nelson for his financial support and technical advice. Drs. Frohmann, Quintana, and Wood must also be thanked most heartily.

End Notes

1. Included in the thesis, but not in this paper, are formal informetric modellings of the n-gram distributions, analysis of the discrimination values of the n-

grams, and retrieval performance evaluations.

2. Only one token for each type present in a song was included in the calculations, thus removing the effect of multiple tokens representing the same type (i.e. within-song redundancy). This was done to better understand how the types might be used as individual identifiers of songs.

3. Other values have been cited (see Heaps 1978; Losee 1990) but for the purposes of this study the author felt it best to chose only the Shannon value in an effort to simplify comparisons.

References

Downie, J. Stephen. 1997. "Evaluating a simple approach to music information retrieval: Representing melodies as collections of 'musical words'.@ Ph.D. proposal. University of Western Ontario.

Downie, J. Spehen. 1997a. "Informetrics and music information retrieval." In *Communication and Information in Context: Society, Technology and the Professions: Proceedings of the 25th Annual Conference of the Canadian Association for Information Science, 8-10 June 1997, St. John=s, Newfoundland.* Ed. Bernd Frohmann. Toronto: Canadian Association for Information Science.

Downie, J. Stephen. 1995. "The MusiFind Music Information Retrieval Project, Phase III: Evaluation of indexing options." In *Connectedness: Information, Systems, People, Organizations: Proceedings of the 23rd Annual Conference of the Canadian Association for Information Science, 7-10 June 1995, Edmonton, Alberta.* Ed. Hope A. Olson and Dennis B. Ward. Edmonton, Alberta: School of Library and Information Studies, University of Alberta.

Downie, J. Stephen. 1994. "The MusiFind Musical Information Retrieval Project, Phase II: User assessment survey." In *The Information Inudstry in Transition: Proceedings of the 22nd Annual Conference of the Canadian Association for Information Science, 25-27 May 1994 Montreal, Quebec.* Toronto: Canadian Association for Information Science.

Dowling, W.J. 1978. "Scale and contour: two components of a theory of memory for melodies." *Physocological Review* 85: 341-54.

Ghias, A., J. Logan, D. Chamberlin, and B.C. Smith. 1995. "Query by humming." *Proceedings of the ACM Multimedia 95.* San Francisco, November.

Heaps. H.S. 1978. *Information Retrieval: Computational and Theoretical Aspects.* New York: Academic Press.

Huron, David. 1991. "Humdrum: Music tools for UNIX systems." *Computing in Musicology* 7:66-67.

Keller, Kate Van Winkle and Carolyn Rabsen. 1980. *National Tune Index, 18th Century Secular Music.* New York: University Music Edition.

Kinnucan, Mark T. 1996. Private meeting held at Graduate School of Library and Information Science, University of Western Ontario.

Losee, Robert M. 1990. *The Science of Information.* San Diego: Academic Press.

McNab, Roger J., Lloyd A. Smith, Ian H. Whitten, Clare Henderson, and Sally Jo Cunningham. 1996. "Towards the digital music library: Tune retrieval from acoustic input." In *Digital Libraries 96: Proceedings of the ACM Digital Libraries Conference. Bethesda, Maryland.*

RISM. 1997. *Handbook: Repertoire International des Source Musicales: series A/H: Music Manuscripts after 1600.* Munich: K.G. Saur.

Shannon, C.E. and W. Weaver. 1949. *The Mathematical Theory of Communication.* Urbana, Ill.: University of Illinois Press.

Shannon, C.E. 1951. *Prediction and entropy of printed English.* Bell System Technical Journal 30: 50-65.

Tague-Sutcliffe, J., J.S. Downie and Shane Dunne. 1993. "Name that tune: An introduction to musical information retrieval." In *Information as a Global Commodity: Communication, Processing and Use: Proceedings of the 21st Annual Conference of the Canadian Association for Information Science, 12-14 July 1993, Antigonish, Nova Scotia.* Toronto: Canadian Association for Information Science.