A qualitative assessment of measures for the evaluation of a cover song identification system

Joan Serrà. Music Technology Group, Universitat Pompeu Fabra, Barcelona, Spain. (jserra@iua.upf.edu)

ISMIR 2007
Vienna, Austria

Music Technology Group

Pharos

ABSTRACT

"*The evaluation of effectiveness in Information Retrieval systems has been developed in parallel to its evolution, generating a great amount of proposals to achieve this process. This paper focuses on a particular task of Music Information Retrieval: a system for Cover Song Identification. We present a concrete example and then try to elucidate which metrics work best to evaluate such a system. We end up with two evaluation measures suitable for this problem: bpref and Normalized Lift Curves.*"

EVALUATION MEASURES

- False / True Positives and Negatives (TP, FP, TN, FN)
- Sensitivity and Specificity
- Fallout Rate
- Receiver Operating Characteristic (ROC) curve
- Lift Curve
- Precision and Recall
- Precision-Recall curve
- Break-even point
- F-measure
- Average Precision (AP)
- Reciprocal Rank (RR)
- Discounted Cumulative Gain (DCG)
- Binary Preference-based measure (bpref / bpref-10)

CASE STUDY: COVER SONG IDENTIFICATION SYSTEM

Main characteristics of the system:

- We have a database of 2054 songs (|D| = 2054), labelled into 451 different groups (or ``canonical'' song versions).
- The average number of covers per song is 4.24, ranging from 1 (the original song + 1 cover) to 14.
- The length of the answer set is set to 14 in order to be able to present to a potential user all the relevant songs in a single output list.

Test Framework

➔ We manually annotate and rank several synthetic sets of prototypical answers to different queries in order to try to elucidate which measure best fits our criteria.
➔ For a set of queries $S_q=\{q_1,...,q_{Nq}\}$, we define a set of answer sets $S_a=\{A_1,...,A_{Nq}\}$, where each $A_k=\{a_{k,1},a_{k,2},...,a_{k,14}\}$.
➔ We intentionally rank the answers A_k from most to least important for us. This is the way we define the relevance of the answer sets. This also helps to observe which measures are more suitable.

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}	a_{11}	a_{12}	a_{13}	a_{14}	$\|R_q\|$
$q_1 \Rightarrow A_1$				⋆											1
$q_2 \Rightarrow A_2$	⋆	⋆	⋆		⋆										7
$q_3 \Rightarrow A_3$						⋆	⋆	⋆		⋆					7
$q_4 \Rightarrow A_4$		⋆		⋆		⋆		⋆							14
$q_5 \Rightarrow A_5$	⋆					⋆	⋆	⋆							14
$q_6 \Rightarrow A_6$															4

Table 1. Test answer set example. It consists of 6 manually labelled answer sets (A_i) answering 6 hypothetical queries (q_i). These answer sets are composed of 14 ranked documents ($A_i = \{a_1, \ldots, a_{14}\}$), and they are ordered from most valuable (A_1) to less valuable (A_6). The "⋆" symbol in (i, j) cell denotes that the a_j document is relevant for the i-th query. Last column ($|R_q|$) denotes the total number of covers for the query q_i that can be found in the database.

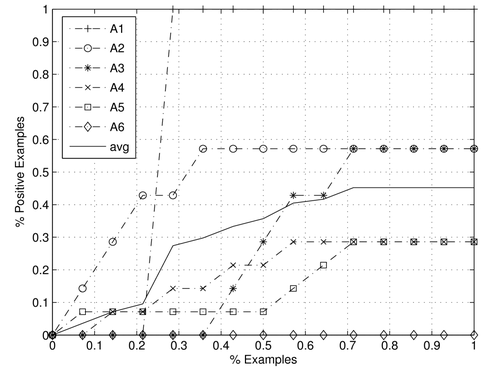
Evaluation measures for a Cover Song Identification system

Measure	A_1	A_2	A_3	A_4	A_5	A_6
TP	1	4	4	4	4	0
FP	13	10	10	10	10	14
FN	0	3	3	10	10	14
TN	2040	2037	2037	2030	2030	2036
Accuracy	0.994	0.994	0.994	0.990	0.990	0.991
Sensitivity	1.000	0.571	0.571	0.285	0.285	0.000
Specificity	1.000	0.998	0.998	0.995	0.995	0.998
Fallout rate	0.006	0.005	0.005	0.005	0.005	0.007
Precision	0.071	0.286	0.286	0.286	0.286	0.000
Recall	1.000	0.571	0.571	0.286	0.286	0.000
Break-even point	0.3	0.7	0.3	0.5	0.4	0.1
AP	0.250	0.950	0.307	0.500	0.496	0.000
F-measure	0.133	0.381	0.381	0.286	0.286	0.000
RR	0.018	0.145	0.038	0.074	0.095	0.000
DCG	0.721	3.974	1.987	3.203	2.371	0.000
bpref	-2.000	0.550	0.143	0.235	0.194	0.000
bpref-10	0.727	0.563	0.395	0.256	0.232	0.000
bpref*	0.800	0.564	0.428	0.260	0.239	0.000

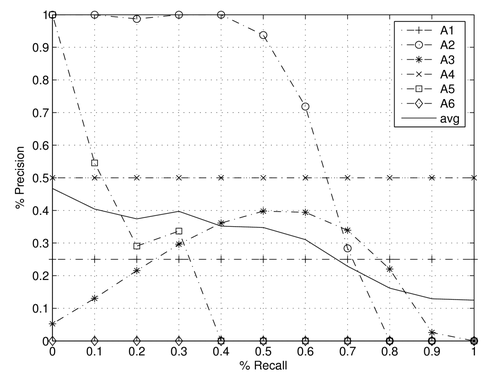
Table 2. Results for different measures for the test case example shown in table 1. The columns correspond to the value of the evaluation measure for the answer set A_i in the forementioned example set.

✗ TP, FP, TN, FN: Do not consider the rank of correctly classified items nor the toal number of relevant documents per query.

✗ Accuracy, Specificity, Fallout rate, ROC and Lift curves: The same as before + Skew of data (99.9% of the documents in the not relevant category).
✓ Useful variant = **Normalized Lift curves**.



✗ Precision and Recall: Do not take the position of correctly classified items into account. Recall better than Precision.
✗ F-measure and others combining Precision and Recall: Same drawbacks as these two.
✗ Precision-Recall curve: Does not measure if we have retrieved all possible elements. Problems in interpolation. Sometimes difficult to interpret.



✗ AP, RR and DCG: Ranking matters a lot.
✓ **Bpref, bpref-10, bpref***: Seems to work well for practically all the answer sets tested.

References

• R. Baeza-Yates and B. Ribeiro Neto. Modern Information Retrieval. ACM Press Books, 1999.
• C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. SIGIR'04, (27), 2004.
• C. D. Manning, R. Prabhakar and H. Schutze. An introduction to Information Retrieval. Cambridge University Press, Cambridge, England, preliminary draft ed., 2007. Online version at http://www.informationretrieval.org
• E. M. Voorhees and L. P. Buckland. Common evaluation measures. in Proc. of Text Retrieval Conference, 2006. Appendix.
• N. Ye. The handbook of Data Mining. Lawrence Erlbaum Associates, 2003.