Répertoire International de Littérature Musicale
International Repertory of Music Literature
Internationales Repertorium der Musikliteratur

A Comparison between Users' free-text queries and RILM index terms

Shuheng Wu
Queens College, CUNY
Yun F. Henshaw
RILM

Répertoire International de Littérature Musicale
International Repertory of Music Literature
Internationales Repertorium der Musikliteratur

Agenda

- Introduction
- Research questions
- Study design
- Data analysis results & suggestions
- Limitations & next steps

Why do we study user queries

Problems exist in controlled vocabularies

- rigid in structure
- unfamiliar to users
- artificial and biased
- indexing inconsistency
- expensive and time-consuming to develop & maintain
- slow to adopt new terminology

Why do we study user queries

- NISO recommendation
- Validated by prior research
- No previous studies on search-logs of music literature databases

Research Questions

1. What are the **categories** of users' free-text queries?
2. How do users' free-text queries differ from controlled vocabularies?

Répertoire International de Littérature Musicale
International Repertory of Music Literature
Internationales Repertorium der Musikliteratur

Study Design

- **Data source**
  RILM Abstracts of Music Literature
- **Data collection**
  - 7,924 unique user queries from December 2015
  - random sample: 367 user queries
  - three types of mapping: perfect match, partial match, no match

Répertoire International de Littérature Musicale
International Repertory of Music Literature
Internationales Repertorium der Musikliteratur

Dataset

invalid & non-user created queries

- Sample: 367 − 28 = 339 queries (valid, user-created)

178 (52.51%)

single-word queries

e.g., zydeco, apartheid, Atlantis

single-concept queries

e.g., music festivals,
murky bass,
Bach, Johann Bernhard

161 (47.49%)

multi-concept queries

e.g., army and ww2
Korea and censorship
Hindemith double bass
Bastista Cuba

350

user-created search terms

7

11 categories of User-created Queries

Category	Single-word/concept queries		Multi-concept queries	
	Count of terms	% of the group	Count of terms	% of the group
Personal name	67	37.64%	126	36.00%
Work title	63	35.39%	69	19.71%
Topical term	35	19.66%	116	33.14%
Geographic name	4	2.25%	12	3.43%
Corporate body name	4	2.25%	5	1.43%
Instrument	3	1.69%	10	2.86%
Identifier	2	1.12%	0	0.00%
Chronological term	0	0.00%	7	2.00%
Format	0	0.00%	1	0.29%
Language	0	0.00%	1	0.29%
Document type	0	0.00%	3	0.86%
Total	178	100%	350	100%

Sub-categories of Work Titles

Subcategory	Single-word/concept queries		Multi-concept queries	
	Count of terms	% of the group	Count of terms	% of the group
Musical work title	19	30.16%	46	66.67%
Music book title	18	28.57%	8	11.59%
Journal article title	11	17.46%	9	13.04%
Film or music video title	5	7.94%	2	2.90%
Literature work title	3	4.76%	1	1.45%
Journal title	2	3.17%	0	0.00%
Cannot be determined	5	7.94%	3	4.35%
Total	63	100%	69	100%

Categories of User-created Queries

Category	Count of terms	% of the group
Personal name	193	36.55%
Topical term	151	28.60%
Work title	**132**	**25%**
Geographic name	16	3.03%
Instrument	13	2.46%
Corporate body name	9	1.70%
Chronological term	7	1.33%
Document type	3	0.57%
Identifier	2	0.33%
Format	1	0.19%
Language	1	0.19%
Total	528	100%

- RILM headwords:
  - Personal name
  - Topical term
  - Geographical name
  - Instrument

Suggestion:
RILM could consider instating musical work titles (e.g., titles that are well known and those without known composers) as a type of headword.

Comparison to RILM's Index Terms

- Single-word/concept queries:

  178 − 22 = 156 user-created search terms for comparison

author, title, & identifier field searches

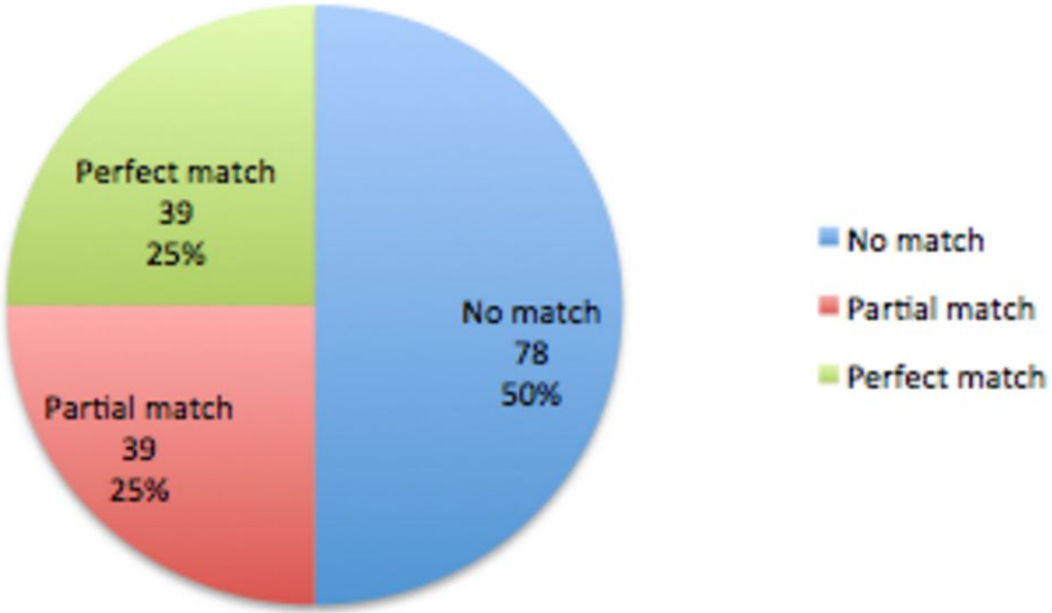
- Multi-concept queries:

  350 − 20 = 330 user-created search terms for comparison

author, title, document type field searches

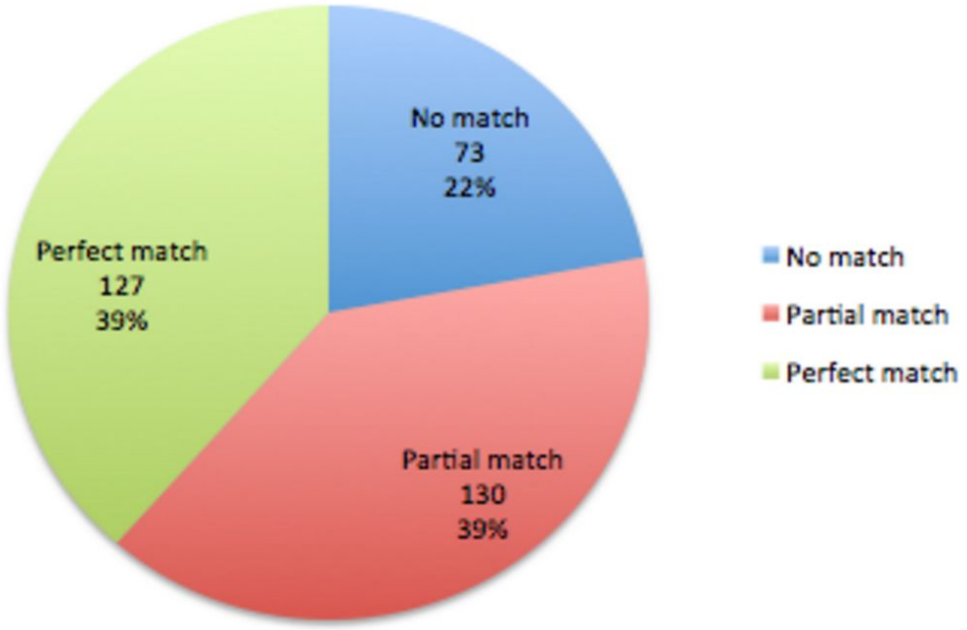
Comparison to RILM's Index Terms

- Single-word/concept queries



Comparison to RILM's Index Terms

- Multi-concept queries



Comparison to RILM's Index Terms

	Perfect match	partial match	no match	Total
Count of terms	166	169	151	486
% of the group	34.16%	34.77%	31.07%	100%

Compared to the findings of previous studies, RILM's index terms align well with users' search terms:

- Mering's (2003) study: 40%-50% of user queries did not match
- Westterstrom's (2008) study: 75% of user tags did not match

Comparison to RILM's Index Terms by Categories

- Single-word/concept queries

category	Perfect match		Partial match		No match	
	Counts of terms	% of the group	Counts of terms	% of the group	Counts of terms	% of the group
Personal name	15	38.46%	22	52.38%	20	26.67%
Work title	5	12.82%	8	19.05%	40	53.33%
Topical term	15	38.46%	7	16.67%	13	17.33%
Geographic name	2	5.13%	1	2.38%	1	1.33%
Corporate body name	2	5.13%	1	2.38%	1	1.33%
Instrument	0	0.00%	3	7.14%	0	0.00%
Total	39	100.00%	42	100.00%	75	100.00%

Comparison to RILM's Index Terms by Categories

- Multi-concept queries

category	Perfect match		Partial match		No match	
	Counts of terms	% of the group	Counts of terms	% of the group	Counts of terms	% of the group
Personal name	17	13.39%	83	62.88%	17	23.94%
Work title	10	7.78%	23	17.42%	26	36.62%
Topical term	76	59.84%	17	12.88%	23	32.39%
Corporate body name	1	0.79%	2	1.52%	2	2.28%
Geographic name	10	7.87%	1	0.76%	1	1.41%
Instrument	10	0.79%	0	0.00%	0	0.00%
Chronological term	1	79.00%	6	4.55%	0	0.00%
Document type	1	79.00%	0	0.00%	2	2.82%
Format	1	79.00%	0	0.00%	0	0.00%
Total	127	100.00%	132	100.00%	71	100.00%

Comparison to RILM's Index Terms by Categories

- Multi-concept queries

category	Perfect match		Partial match		No match	
	Counts of terms	% of the group	Counts of terms	% of the group	Counts of terms	% of the group
Per		13.39%	83	62.88%	17	23.94%
		.78%	23	17.42%	26	36.62%
		9.84%	17	12.88%	23	32.39%
name	1	0.79%	2	1.52%	2	2.28%
Geographic na	10	7.87%	1	0.76%	1	1.41%
Instrument	10	0.79%	0	0.00%	0	0.00%
Chronological term	1	79.00%	6	4.55%	0	0.00%
Document type	1	79.00%	0	0.00%	2	2.82%
Format	1	79.00%	0	0.00%	0	0.00%
Total	127	100.00%	132	100.00%	71	100.00%

Suggestion:

Integrate into RILM's thesaurus some significant chronological terms in music history (e.g., 1800-1850, 19th century).

Reasons for partial-match search terms

Category	Reason for partial match	Count of terms	Suggested solution
Personal name	partial name	102	name suggestion, name disambiguation, thesauru-based autocompletion
	misspelling	2	spell-checker
	alternative name	1	inclusion in the authority file
Work title	incomplete title	26	title suggestion, title disambiguation, thesaurus-based autocompletion, and adoptin LC's uniform titles
	related title	2	ontology
	alternative title	1	inclusion in the authority file
	variant spelling	1	steeming
	singular usage	1	title suggestion

Category	Reason for partial match	Count of terms	Suggested solution
Topical term	singular usage	11	stemming
	broader term	4	inclusion in the thesaurus, linking index terms with hierarchical relationship
	narrower term	2	inclusion in the thesaurus, linking index terms with hierarchical relationship
	synonymous term	1	inclusion in the thesaurus
	variant spelling	3	stemming
	phrase in different word order	2	inclusion in the thesaurus
	the use of an abbreviation	1	inclusion in the thesaurus
Chronological term	the use of fuller form	5	inclusion in the thesaurus
	the use of an acronym	1	inclusion in the thesaurus
Corporate body name	alternative name	2	inclusion in the authority file
	partial name	1	name suggestion, name disambiguation, thesauru-based autocompletion
Instrument	related term	2	linking index terms with associative relationship
	broader term	1	linking index terms with hierarchical relationship
Geographic name	alternative name	1	inclusion in the authority file
	misspelling	1	spell-checker

Reasons for no-match search terms

Category	Reason for partial match	Count of terms	Suggested solution
Work title	not indexed as subject terms in RILM	64	inclusion in the authority file
	use of a different language	2	inclusion in the authority file
Personal name	not indexed as subject terms in RILM	35	inclusion in the authority file
	misspelling	1	spell-checker
	use of a different language	1	inclusion in the authority file
Corporate body name	not indexed as subject terms in RILM	3	inclusion in the authority file
Geographic name	historical place name not indexed in RILM	1	inclusion in the authority file
	alternative name	1	inclusion in the authority file
Document type	not indexed as subject terms in RILM	2	

Reasons for no-match search terms

Category	Reason for partial match	Count of terms	Suggested solution
Work title	not indexed as subject terms in RILM	64	inclusion in the authority file
	use of a different language	2	inclusion in the authority file
Personal nam	not indexed as subject terms in RILM	35	in
		1	
		1	
Corporate bod name	RILM	3	
Geographic name	historical place name not indexed in RILM	1	inclusion in the authority file
	alternative name	1	inclusion in the authority file
Document type	not indexed as subject terms in RILM	2	

"Santa Elena," the capital of Spanish Florida from 1566 to 1587. RILM users may have the need for cultural music information in temporal nature.

Suggestion:
Aggregate historical terms to RILM's controlled vocabularies, and associate them with contemporary terms representing the same concepts or entities.

Category	Reason for partial match	Count of terms	Suggested solution
Topical term	non-music term	10	ontology
	new or popular term	6	inclusion in the thesaurus
	broader term		linking index terms
	narrower term		index terms
	related term	2	inclusion ... saurus, linking index terms with associat... elationship
	use of a different language	4	**multilingual thesaurus**
	misspelling	3	spell-checker
	invalid term	3	

Suggestion:
Adopt a collaborative approach to developing or enhancing multilingual controlled vocabularies with other cultural heritage institutions in different countries and music literature databases.

Limitations

- Search log analysis is an unobtrusive research method
- No differentiation between initial queries and modified queries
- Search-log data cannot provide demographic information of the users

Nest steps

- Extract and identify relationships among co-occur user-created search terms
- Evaluate RILM's controlled vocabularies by comparing the number of search hits with and without index strings in the bibliographic records

Répertoire International de Littérature Musicale
International Repertory of Music Literature
Internationales Repertorium der Musikliteratur

Thank you!