

**Enabling Entity Retrieval by Exploiting Wikipedia
as a Semantic Knowledge Source**

A Thesis

Submitted to the Faculty

of

Drexel University

by

Sofia Jeon

in partial fulfillment of the

requirements for the degree

of

Doctor of Philosophy

August 2011

© Copyright 2011

Sofia Jeon. All rights reserved.

Dedications

To Dr. Giorgio Ingargiola

and

To My Mother and Father

Acknowledgments

I hereby acknowledge the fact that I have been awarded a 2011 Eugene Garfield Doctoral Dissertation Fellowship by Beta Phi Mu the International Library & Information Studies Honor Society in recognition of the work described in this dissertation.

I thank the members of my dissertation committee including Dr. Xia Lin, chair and main advisor, Dr. Il-Yeol Song, Dr. Eileen Abels, and Dr. Andrea Forte, all at Drexel University, and Dr. John E. Hopcroft at Cornell University, for their advice and assistance toward a successful completion of this dissertation research.

Table of Contents

LIST OF TABLES	vi
LIST OF FIGURES	x
ABSTRACT.....	xvi
1. INTRODUCTION	1
2. BACKGROUND AND MOTIVATION	3
3. RESEARCH PROBLEM, GOAL, AND QUESTIONS.....	5
4. RELATED WORK.....	7
4.1 Entity Search, Retrieval, and Ranking.....	7
4.2 Information Extraction from Wikipedia	29
4.3 Information Retrieval on Wikipedia Data.....	60
5. CONCEPTUALIZATION.....	72
5.1 Conceptual Stance	72
5.2 Ontological Scheme.....	73
5.3 Semantic Entity Typing.....	80
6. IMPLEMENTATION: INFORMATION EXTRACTION	82
6.1 Information Extraction Process	82
6.1.1 Information Source	82
6.1.2 Direct Extraction of Information	84
6.1.3 Indirect Derivation of Information	104
6.2 Information Storage and Organization	113
6.3 Information Extraction Statistics	118

7.	IMPLEMENTATION: INTERFACE CONSTRUCTION	119
7.1	Film-Related Interface Examples	119
7.2	Interface Design and Implementation.....	133
7.2.1	Interface Functions	133
7.2.2	Search/Retrieval Process.....	135
7.2.3	Interface Implemented	136
7.3	Interface Usage Example	157
8.	EVALUATION	158
8.1	Evaluation Methodology Overview	158
8.2	Evaluation: Information Extraction	160
8.2.1	Dataset	160
8.2.2	Methods	160
8.2.3	Results.....	162
8.3	Evaluation: Information Retrieval	166
8.3.1	Experimental Design.....	166
8.3.2	Experimental Procedures	167
8.3.3	Experimental Results	172
9.	CONCLUSION.....	188
	LIST OF REFERENCES	192
	APPENDIX: IRB APPROVAL NOTICE	205
	VITA.....	213

List of Tables

1.	Zaragoza et al.: Sample queries and entity relevance judgments (taken from [Zar07]).....	15
2.	INEX 2008 XER Track: Query topics and entity types in testing set.....	20
3.	Craswell et al.: Examples of entity attributes extracted (taken from [Cra09]).....	25
4.	Craswell et al.: Examples of entity relations extracted (taken from [Cra09]).....	26
5.	YAGO: Number of facts as of 2007 (taken from [Suc07]).....	38
6.	YAGO: Accuracy of facts as of 2007 (taken from [Suc07]).....	38
7.	YAGO: Number of entities as of 2008 (taken from [Suc08]).....	40
8.	YAGO: Number of facts involving largest relations as of 2008 (taken from [Suc08])	40
9.	YAGO: Sample facts (taken from [Suc07]).....	40
10.	YAGO: Sample queries (taken from [Suc07]).....	42
11.	DBpedia: Examples of typed literals (taken from [Auer07a])	51
12.	DBpedia: Level-1 classes in ontology.....	52
13.	DBpedia: Component datasets	53
14.	DBpedia: Current statistics on dataset content.....	53
15.	LinkedMDB: Overall statistics on dataset content (taken from [Has09]).....	57
16.	LinkedMDB: Statistics on sample entity types (taken from [Has09])	57
17.	LinkedMDB: Statistics on interlinked resources (taken from [Has09]).....	57
18.	LinkedMDB vs. PanAnthropon: Comparison of number of entities	59
19.	PanAnthropon: Film-domain-oriented ontology	74
20.	DBpedia: Upper-level classes in ontology	76
21.	PanAnthropon: Entity types/subtypes	80
22.	PanAnthropon: Simplified entity types/subtypes.....	81
23.	PanAnthropon: Cue words/phrases for extraction of also_known_as facts	85

24. PanAnthropon: Film attributes extracted from infobox	90
25. PanAnthropon: Taxonomy of super-categories.....	92
26. PanAnthropon: Variations of film cast section titles considered	97
27. PanAnthropon: Types of film-centric facts extracted	101
28. PanAnthropon: Types of film-award-related facts extracted	103
29. PanAnthropon: Inverse attributes derived from infobox-based attributes	105
30. PanAnthropon: Types of facts derived from film-cast-related facts	107
31. PanAnthropon: Types of facts derived about temporal/geographical inclusion relations....	107
32. PanAnthropon: Classes derived from super-categories	109
33. PanAnthropon: Types of facts derived from category-related facts.....	110
34. PanAnthropon: Types of facts derived from film-award-related facts.....	111
35. PanAnthropon: Types of person-name-related facts derived.....	112
36. PanAnthropon: Structure of database table <code>Class</code>	115
37. PanAnthropon: Structure of database table <code>Page</code>	115
38. PanAnthropon: Structure of database table <code>Entity</code>	115
39. PanAnthropon: Structure of database table <code>Entity_Fact</code>	116
40. PanAnthropon: Structure of database table <code>Category_Super</code>	116
41. PanAnthropon: Structure of database table <code>Category</code>	116
42. PanAnthropon: Structure of database table <code>Category_Fact</code>	117
43. PanAnthropon: Structure of database table <code>Attribute</code>	117
44. PanAnthropon: Structure of database table <code>Attribute_Fact</code>	117
45. PanAnthropon: Overall information extraction statistics.....	118
46. PanAnthropon: Number of entities per entity type	118
47. PanAnthropon: Summary results of info extraction evaluation: number of facts	163
48. PanAnthropon: Summary results of info extraction evaluation: precision/recall	163

49. PanAnthropon: Summary results of info extraction evaluation: average precision/recall ...	163
50. PanAnthropon: Detailed results of info extraction evaluation	164
51. PanAnthropon: Experimental task codes and corresponding question orderings	170
52. PanAnthropon: Pre-experimental-task questionnaire responses: student type.....	172
53. PanAnthropon: Pre-experimental-task questionnaire responses: student major	173
54. PanAnthropon: Pre-experimental-task questionnaire responses: age	173
55. PanAnthropon: Pre-experimental-task questionnaire responses: info search experience	173
56. PanAnthropon: Pre-experimental-task questionnaire responses: info search frequency	173
57. PanAnthropon: Experimental task results: per-group avg/max/min precision/recall.....	177
58. PanAnthropon: Experimental task results: per-subject average precision/recall #1	177
59. PanAnthropon: Experimental task results: per-subject average precision/recall #2	177
60. PanAnthropon: Experimental task results: per-subject average precision/recall #3	178
61. PanAnthropon: Detailed experimental task results	178
62. PanAnthropon: Post-experimental-task questionnaire responses #1	179
63. PanAnthropon: Post-experimental-task questionnaire responses #2	179
64. PanAnthropon: Reasons for effectiveness of PanAnthropon interface: design	180
65. PanAnthropon: Reasons for effectiveness of PanAnthropon interface: function.....	180
66. PanAnthropon: Yes ($N=1$) response on effectiveness of IMDb interface.....	181
67. PanAnthropon: Maybe ($N=7$) responses on effectiveness of IMDb interface	181
68. PanAnthropon: No ($N=25$) responses on effectiveness of IMDb interface.....	182
69. PanAnthropon: Yes ($N=32$) responses on effectiveness of PanAnthropon interface.....	183
70. PanAnthropon: Maybe ($N=1$) response on effectiveness of PanAnthropon interface	184
71. PanAnthropon: Responses ($N=33$) on relative effectiveness of PanAnthropon vs. IMDb ..	184
72. PanAnthropon: Maybe ($N=2$) responses on usability of PanAnthropon interface.....	186
73. PanAnthropon: No ($N=2$) responses on interest in using similar interfaces	186

74. PanAnthropon: Maybe ($N=2$) responses on interest in using similar interfaces	187
75. PanAnthropon: Yes ($N=29$) responses on interest in using similar interfaces	187

List of Figures

1.	Cheng et al.: Entity search system architecture (taken from [Che07a]).....	11
2.	INEX 2008 XER Track: Sample query topic in testing set.....	19
3.	Craswell et al.: Pattern matching rules for attribute extraction (taken from [Cra09])	24
4.	YAGO: Literal classes in data model (taken from [Suc08]).....	33
5.	YAGO: Axiomatic rules in data model (taken from [Suc07])	35
6.	YAGO: Sample facts in raw text format (taken from YAGO project site).....	40
7.	YAGO: Snippet of RDFS version (taken from YAGO project site).....	41
8.	YAGO: Snippet of subsumption hierarchy (taken from YAGO project site).....	41
9.	DBpedia: Overview of components (taken from [Auer07b])	48
10.	Wikipedia: MediaWiki markup of infobox template and its output (taken from [Auer07a])	49
11.	DBpedia: Snippet of ontology (taken from DBpedia project site).....	52
12.	DBpedia: Sample resource accessed via regular Web browser	54
13.	DBpedia: Interlinked data resources (taken from Linked Data site).....	54
14.	LinkedMDB: Sample entities (taken from [Has09]).....	56
15.	LinkedMDB: Sample resource accessed via regular Web browser	58
16.	Wikipedia: Sample search result.....	61
17.	Wikipedia: Filmography section of page on Werner Herzog	62
18.	Wikipedia: Category page on films directed by Werner Herzog	62
19.	Wikiwix: Search interface.....	63
20.	Wikiwix: Sample search result.....	63
21.	Powerset: Sample search result.....	64
22.	Similpedia: Search interface	65
23.	Similpedia: Sample search result	65

24. WikiWax: Search suggestion — no matching titles found	66
25. WikiWax: Search suggestion — numerous matching titles found	66
26. WikiMindMap: Sample search result.....	67
27. Koru: Search interface with sample search result (taken from [Mil07b])	68
28. YAGO: Query examples (taken from YAGO project site)	69
29. YAGO: Dropdown menu containing predicates	70
30. YAGO: Sample query result	70
31. DBpedia: Query Builder interface (taken from [Auer07a])	71
32. Wikipedia: Sample film release section.....	83
33. Wikipedia: Sample film page abstract section	85
34. Wikipedia: Sample film infobox section.....	87
35. Wikipedia: Sample film categories section.....	91
36. Wikipedia: Sample film cast section #1.....	98
37. Wikipedia: Sample film cast section #2.....	98
38. Wikipedia: Sample film cast section #3.....	98
39. Wikipedia: Sample film cast section #4.....	99
40. Wikipedia: Sample film cast section #5.....	99
41. Wikipedia: Sample film cast section #6.....	99
42. PanAnthropon: Database table schemas based on Hierarchical Tree Model.....	114
43. PanAnthropon: Database table schemas based on Common Relational Model.....	114
44. PanAnthropon: Database table schemas based on Entity–Attribute–Value Model	114
45. IMDb: Homepage	119
46. IMDb: Movie-related menus.....	120
47. IMDb: Now Playing page	120
48. IMDb: Basic search menu.....	121

49. IMDb: Sample search-by-title result	121
50. IMDb: Advanced Search page	123
51. IMDb: Part of Advanced Title Search interface #1.....	124
52. IMDb: Part of Advanced Title Search interface #2.....	125
53. IMDb: Part of Advanced Title Search interface #3.....	126
54. IMDb: Part of Advanced Title Search interface #4.....	127
55. IMDb: Sample Advanced Title Search result	128
56. IMDb: Options for browsing titles/people	129
57. IMDb: Interface for browsing titles by genre.....	129
58. FMDb: Homepage.....	130
59. FMDb: Search suggestion menu	130
60. FMDb: Sample search result.....	130
61. LinkedMDB: Homepage.....	131
62. LinkedMDB: Interface for browsing films	131
63. Netflix: Main menu bar.....	132
64. Netflix: Search suggestion menu	132
65. Netflix: Recommendation functionality.....	132
66. PanAnthropon: Flowchart of search process using interface	135
67: PanAnthropon: Homepage	136
68. PanAnthropon: GERQ function interface initial screen.....	137
69. PanAnthropon: GERQ function initial query form.....	137
70. PanAnthropon: GERQ function entity type selection menu	137
71. PanAnthropon: GERQ function query form after entity type selection.....	138
72. PanAnthropon: GERQ function entity subtype selection menu.....	138
73. PanAnthropon: GERQ function simplified entity subtype selection menu	138

74. PanAnthropon: GERQ function query form after entity subtype selection	139
75. PanAnthropon: GERQ function attribute selection menu	139
76. PanAnthropon: GERQ function query form after attribute selection #1.....	140
77. PanAnthropon: GERQ function query form after value text input	140
78. PanAnthropon: GERQ function value selection menu	140
79. PanAnthropon: GERQ function query form after attribute selection #2.....	141
80. PanAnthropon: GERQ function query form after value selection	141
81. PanAnthropon: GERQ function query processing	141
82. PanAnthropon: GERQ function query result	142
83. PanAnthropon: GERQ function entity fact window	143
84. PanAnthropon: GERQ function entity Wikipedia page window	143
85. PanAnthropon: SECQ function interface initial screen	144
86. PanAnthropon: SECQ function initial query form.....	144
87. PanAnthropon: SECQ function query form after entity subtype selection #1	144
88. PanAnthropon: SECQ function query form after entity subtype selection #2	144
89. PanAnthropon: SECQ function query form after entity name selection.....	145
90. PanAnthropon: SECQ function query result.....	145
91. PanAnthropon: ECFQ function interface initial screen	146
92. PanAnthropon: ECFQ function initial query form.....	146
93. PanAnthropon: ECFQ function query form after entity subtype selection	146
94. PanAnthropon: ECFQ function query form after entity 2 name selection.....	147
95. PanAnthropon: ECFQ function query result.....	147
96. PanAnthropon: DRFQ function interface initial screen	148
97. PanAnthropon: DRFQ function initial query form	148
98. PanAnthropon: DRFQ function query form after entity 1 name selection	148

99. PanAnthropon: DRFQ function query form after entity 2 name selection	149
100. PanAnthropon: DRFQ function query result	149
101. PanAnthropon: IRFQ function interface initial screen	150
102. PanAnthropon: IRFQ function initial query form.....	150
103. PanAnthropon: IRFQ function query result #1	150
104. PanAnthropon: IRFQ function query result #2.....	151
105. PanAnthropon: IRFQ function query result #3.....	151
106. PanAnthropon: CBEB function interface initial screen	152
107. PanAnthropon: CBEB function initial query form	152
108. PanAnthropon: CBEB function menu for top super-category selection	152
109. PanAnthropon: CBEB function query form after top super-category selection	153
110. PanAnthropon: CBEB function menu for sub-super-category selection	153
111. PanAnthropon: CBEB function query form at leaf category selection stage.....	153
112. PanAnthropon: CBEB function menu for leaf category selection	154
113. PanAnthropon: CBEB function query processing	154
114. PanAnthropon: CBEB function query result.....	154
115. PanAnthropon: Slide function interface initial screen	155
116. PanAnthropon: Slide function film list retrieval.....	155
117. PanAnthropon: Slide function query form after film list retrieval.....	155
118. PanAnthropon: Slide function menu for film name selection.....	156
119. PanAnthropon: Slide function film info slide	156
120. PanAnthropon: GERQ function query form for sample query	157
121. PanAnthropon: GERQ function sample query result.....	157
122. PanAnthropon: Equations for evaluation on information extraction	158
123. PanAnthropon: Equations for evaluation on information retrieval	159

124. PanAnthropon: How-To-Use page.....	168
125. PanAnthropon: Pre-experimental-task questionnaire.....	169
126. PanAnthropon: Experimental task question set	171
127. PanAnthropon: Post-experimental-task questionnaire	172
128. PanAnthropon: Experimental-task answer set #1	174
129. PanAnthropon: Experimental-task answer set #2	175
130. PanAnthropon: Experimental-task answer set #3	176

Abstract

Enabling Entity Retrieval by Exploiting Wikipedia as a Semantic Knowledge Source

Sofia Jeon, Ph.D.

Xia Lin, Ph.D.

This dissertation research, PanAnthropon FilmWorld, aims to demonstrate direct retrieval of entities and related facts by exploiting Wikipedia as a semantic knowledge source, with the film domain as its proof-of-concept domain of application. To this end, a semantic knowledge base concerning the film domain has been constructed with the data extracted/derived from 10,640 Wikipedia pages on films and additional pages on film awards. The knowledge base currently contains 209,266 entities and 2,345,931 entity-centric facts. Both the knowledge base and the corresponding semantic search interface are based on the coherent classification of entities. Entity-centric facts are also consistently represented as <entity, attribute, value, note> tuples. The semantic search interface (<http://dlib.ischool.drexel.edu:8080/sofia/PA/>) supports multiple types of semantic search functions, which go beyond the traditional keyword-based search function, including the main General Entity Retrieval Query (GERQ) function, which is concerned with retrieving all entities that match the specified entity type, subtype, and semantic conditions and thus corresponds to the main research problem. Two types of evaluation have been performed in order to evaluate (1) the quality of information extraction and (2) the effectiveness of information retrieval using the semantic interface. The first type of evaluation has been performed by inspecting 11,495 film-centric facts concerning 100 films. The results have confirmed high data quality with 99.96% average precision and 99.84% average recall. The second type of evaluation has been performed by conducting an experiment with human subjects. The experiment involved having the subjects perform a retrieval task by using both the PanAnthropon interface and the Internet Movie Database (IMDb) interface and comparing their task performance between the two interfaces. The results have confirmed higher effectiveness of the PanAnthropon interface vs. the

IMDb interface (83.11% vs. 40.78% average precision; 83.55% vs. 40.26% average recall). Moreover, the subjects' responses to the post-task questionnaire indicate that the subjects found the PanAnthropon interface to be highly usable and easily understandable as well as highly effective. The main contribution from this research therefore consists in achieving the set research goal, namely, demonstrating the utility and feasibility of semantics-based direct entity retrieval.

CHAPTER 1: INTRODUCTION

Since its inception in 2001 as a freely-editable collaborative Web encyclopedia project, Wikipedia (<http://www.wikipedia.org/>) has grown rapidly to become one of the most prominent information resources on the Web. As of the time of this writing in July 2011, the English version of Wikipedia (<http://en.wikipedia.org/>) contains close to 3.7 million articles on a wide variety of topics. Corresponding to the rapid, exponential growth of the size of Wikipedia, recent years have witnessed an increasing number of computer/information science researchers working on various projects concerned with Wikipedia. In particular, researchers working in the fields of natural language processing, text mining, information extraction, question answering, etc. have explored various ways to exploit the vast amount of lexical, semantic, and encyclopedic knowledge contained in Wikipedia. In addition, some Semantic Web researchers have turned to Wikipedia for clues to resolving the knowledge acquisition bottleneck due to the scarcity of structured or semi-structured semantic data available on the Web.

The main objective of the research described in this dissertation is to demonstrate the utility and feasibility of a new mode of information retrieval that is concerned with retrieving entities/facts that directly match a given query. It approaches the problem by extracting semantic information from Wikipedia and by deriving/reorganizing entity-centric facts using entity-relevant attributes, based on coherent conceptual schemes of entity/attribute classification and fact representation. For a proof-of-concept application, the research applies the approach to entity/fact retrieval concerning the film domain. The products from the research include a semantic knowledge base containing the entities/facts extracted/derived and a semantic search interface demonstrating the proposed entity/fact retrieval capability. The results of evaluation confirm both the high quality of entity/fact extraction and the high utility of entity/fact retrieval. This dissertation serves as a

comprehensive report on the research involving the project, entitled PanAnthropon FilmWorld (or, PanAnthropon, in short), encompassing the stages of conceptualization, implementation, and evaluation. (Hereafter, “the PanAnthropon project” and “this [dissertation] research” will be used interchangeably, depending on the context.)

The remainder of this dissertation is organized as follows:

- Chapter 2 describes the background and motivation.
- Chapter 3 states the main research problem and goals.
- Chapter 4 discusses related work.
- Chapter 5 states the conceptual position underlying the research.
- Chapter 6 describes the process and results of information extraction.
- Chapter 7 describes the features of the entity search interface.
- Chapter 8 discusses the methods and results of evaluation.
- Chapter 9 concludes the dissertation.

CHAPTER 2: BACKGROUND AND MOTIVATION

Traditional information retrieval is concerned with retrieving documents that are potentially relevant to a user's query. The relevance of a document to a query is usually estimated by lexico-syntactic matching between the terms in the query and those in the document (title). Familiar Web search engines, e.g., Google and Yahoo, only allow the user to express information needs in terms of a query string consisting of one or more keywords, and in response return a list of Web pages that contain all or some of the keywords in the query string, rather than a list of objects of query that directly match information needs. As such, the matching between the query and the query result does not take semantics into account.

The Semantic Web [Ber01] movement aims at transforming the current Web consisting of human-readable, unstructured pages into an intelligent Web of machine-understandable and -processable data. The envisioned transition can also be expressed as one away from the Web of pages (documents) toward the Web of entities ("things" in the broad sense) (cf. OKKAM project (<http://www.okkam.org/>) [Bou07]). What this means is that information retrieval on the Semantic Web is no longer a matter of retrieving documents via semantics-unaware keyword matching but a matter of retrieving entities that satisfy the semantic constraints imposed by the query, i.e., those that are of a specific semantic type and that satisfy the given semantic conditions.

Albeit not in itself a Semantic Web project, Wikipedia has become an important semantic knowledge resource [Zes07b] for various projects involving information extraction, knowledge engineering, and the Semantic Web, due to its unique set of semi-structured semantic features and the huge amount of content covering a wide range of topics. What renders Wikipedia even more interesting is the fact that it can be considered as a self-contained web of entities. Each Wikipedia

article is concerned with one entity, and the given entity is connected with a number of other entities via explicit semantic relations as in infoboxes and wikitables or via implicit semantic relations as in hyperlinks.

Through a pilot study called WikiPhiloSofia (aka The WikiPhil Portal) [Ath09a] [Ath09b] [AthL08a] [AthL08b] [AthL09a] [AthL09b] I demonstrated extracting, retrieving, and visualizing specific facets of information concerning entities of a selected type, namely, philosophers, by exploiting the hyperlinks, categories, infoboxes, and wikitables contained in Wikipedia articles. The Web interface that I created for the study (<http://research.cis.drexel.edu:8080/sofia/WPS/>) enables the user to select a focus of query in the form of an entity (philosopher) or a pair of entities (philosophers) and then to retrieve (philosopher- and non-philosopher-type) entities that satisfy specified relations with respect to the given entity or pair of entities. However, the pilot project did not consider the problem of retrieving entities by type and condition as answers to entity-search queries.

The PanAnthropon project takes up the aforementioned problem left out of the WikiPhiloSofia study [Ath10]. The goal is to enable retrieval of entities that directly answer a given query, given the selection of the type of entities sought and the specification of the conditions to be satisfied by those entities.

CHAPTER 3: RESEARCH PROBLEM, GOAL, AND QUESTIONS

The main research problem that is addressed is how to enable a new mode of information retrieval, namely, entity retrieval, which departs from the traditional framework of word-based, document-centric, indirect information retrieval toward an emerging framework of meaning-based, entity-centric, direct information retrieval. In other words, the problem is about being able to directly retrieve the objects of query rather than being given indirect pointers.

This research addresses the problem by exploiting Wikipedia as a semantic knowledge source. By constructing a semantic knowledge base containing the entities/facts extracted/derived from Wikipedia and by implementing a semantic search interface connected to the knowledge base, the research aims to demonstrate the utility and feasibility of semantics-based retrieval of entities and related facts that directly match the user's information needs.

The execution of this research, therefore, involves both information extraction and information retrieval. Correspondingly, the research questions relevant to this research are as follows:

- (1) What kinds of semantic knowledge can be extracted from Wikipedia?
- (2) What kinds of queries can be answered by using the extracted knowledge?
- (3) How effectively can such queries be answered by using the search interface?

Question 1 concerning the types of semantic knowledge extracted/derived from Wikipedia will be addressed in Chapter 6 where the process and results of information extraction are discussed.

Question 2 concerning the types of queries covered by the semantic search interface will be addressed in Chapter 7 where the design and implementation of the interface are described.

Question 3 concerning the (independent and relative) effectiveness of the search interface will be addressed in Chapter 8 where the methods and results of evaluation are delineated.

CHAPTER 4: RELATED WORK

4.1 Entity Search, Retrieval, and Ranking

Entity search/retrieval/ranking is an emerging field of information retrieval that aims to retrieve/rank entities that match a given query. This dissertation research is concerned with entity search/retrieval, not with ranking. It considers only *exact semantic matching* between a query and entities, i.e., entities must exactly match the selected entity type/subtype and specified semantic conditions to be retrieved as an answer to a query. Such matching is based on the entity-centric facts derived from the semantic information extracted from Wikipedia. Barring the possibility of the existence of a relatively small number of incorrect facts due to the errors in the original information source, it is quite unlikely that the system will retrieve any entities that are not exactly correct answers for a given query, i.e., all entities returned as the result of a query will be *equally* correct. Hence, ranking entities returned as query results, which is appropriate in situations where some query results may be more or less correct than the others, is rather irrelevant to this research, although some kind of query result ordering (of equally correct entities) may be incorporated in the latter. As such, the overview of related works below will not address the details of entity ranking algorithms.

Expert search is an earlier, more restricted form of entity retrieval/ranking. Given a query specifying a field of expertise, the aim of expert search is to produce a ranked list of experts from a list of candidates in a given organization or domain. Examples of earlier systems for expert search include [Mat98] and [Cra01]. Recently, the Enterprise Track of the Text REtrieval Conference (TREC) (<http://trec.nist.gov/>) has since 2005 incorporated the Expert Search task [Cra06] [Sob07] [Bai08], which involves ranking experts given a query topic, commonly by

identifying a set of relevant documents in a collection of corporate documents and then detecting the occurrences of expert names within the former.

The problem of finding and ranking entities (of multiple types) on the Web in general has been studied, for example, by Cheng et al. [Che07a] [Che07b].

Cheng et al. note the fact that what we often search for on the Web are various entities and that the current Web search systems are inadequate for the task. As they see it, the two major barriers to finding entities on the Web concern the fact that the search engines search for information *indirectly* and *individually*. First, the indirectness is concerned with the input/output involved with the current Web search process: Users have to indirectly formulate their information needs for entities as keyword queries. In response, the search engines return a list of Web pages that potentially contain information on the target entities, not the target entities themselves. Second, the individualistic aspect has to do with the matching mechanism used by the current Web search engines: Current Web search engines treat each page individually, despite the fact that information on the target entities may be distributed across multiple pages.

Based on their assessment of the problems in the current Web search mechanism, Cheng et al. devised a conceptual framework, called EntityRank, and a prototype entity search system in order to demonstrate a direct and holistic approach to finding entities on the Web. The data model that constitutes the basis of their framework takes an entity view of the Web, considering the Web as a repository of entities over a collection of documents. The entity search problem is then conceived of in terms of the corresponding input/output formats and matching mechanism based on the entity view.

First, in terms of input, Cheng et al.'s entity search system lets users search for entities directly by specifying target entity types and contextual keywords together in a tuple pattern. From Cheng et al.'s point of view, entity search is essentially *search by context* over the document collection, meaning that the search is done by considering, not only the target entity types, but also the context patterns in which the target entities may appear in documents, which consist of the textual co-occurrences of certain keywords (and other entities). The kinds of context patterns, or the scopes of matching thereof, include `doc` (the same document), `ow` (ordered window), `uw` (unordered window), and `phrase` (exact matching). Another component involved in the query input format is optional content restriction, which specifies restriction on the values/instances matching the target entity type and context pattern.

Example query patterns provided by Cheng et al. [Che07b] are as follows (note: # is used as a prefix to indicate an entity type, while = is used as a content restriction operator to indicate equality):

```
Q1: ow(amazon customer service #phone)
Q2:  (#professor #university #research="database")
Q3: ow(sigmod 2006 #pdf_file #ppt_file)
Q4:  (#title="hamlet" #image #price)
```

Q1 indicates a query looking for (any) Amazon customer service phone numbers. Q2 indicates a query searching for professors whose areas of research concern database and their affiliated universities. Q3 indicates a query searching for PDF files and PPT files of SIGMOD 2006 conference papers that come in both formats. Q4 indicates a query looking for prices and images of the book *Hamlet*.

Second, in terms of output, Cheng et al.'s system is intended to directly provide the entity instances (or literal values) that match a given query. Specifically, the result of a query consists of a ranked list of m -ary entity tuples, each of which is in the form of $t = \langle e_1, e_2, \dots, e_m \rangle$, where each e_i indicates an instance of the entity type E_i sought in the query.

Third, in terms of matching/ranking mechanism, Cheng et al.'s system searches for instances of a specified entity type across all the pages where they occur, so that all matching occurrences will be aggregated to form the final ranking. Cheng et al. rank query results by calculating how well a result entity tuple t appears in the desired query tuple pattern α , across every document d_j in the collection D .

While the computation of query scores is central in Cheng et al.'s EntityRank framework, the details involved are less relevant in comparing their approach/system with those of this dissertation research. Hence I direct the reader to [Che07b] for the details and instead describe Cheng et al.'s system below.

As shown in Figure 1, Cheng et al.'s entity search system architecture consists of offline processing modules (marked by dotted lines) concerned mainly with entity extraction and indexing, which are of interest here, and online processing modules (marked by solid lines) concerned with entity ranking.

Entity extraction: Cheng et al. obtained their Web document collection from the Stanford WebBase Project (<http://diglib.stanford.edu:8091/~testbed/doc2/WebBase/>). They implemented two types of entity extractors, rule-driven and dictionary-driven. The rule-driven extractor tags entities with regular patterns, such as #phone entity and #email entity. The dictionary-driven

extractor is supposedly used for entities whose domains or dictionaries are enumerated. (It is not clear what is meant by “entities whose domains or dictionaries are enumerated”. The examples given by Cheng et al. include #university entity, #professor entity, and #research entity.) Cheng et al. recorded three features for each occurrence of an entity instance: entity instance ID, position (document ID + word offset), and confidence level.

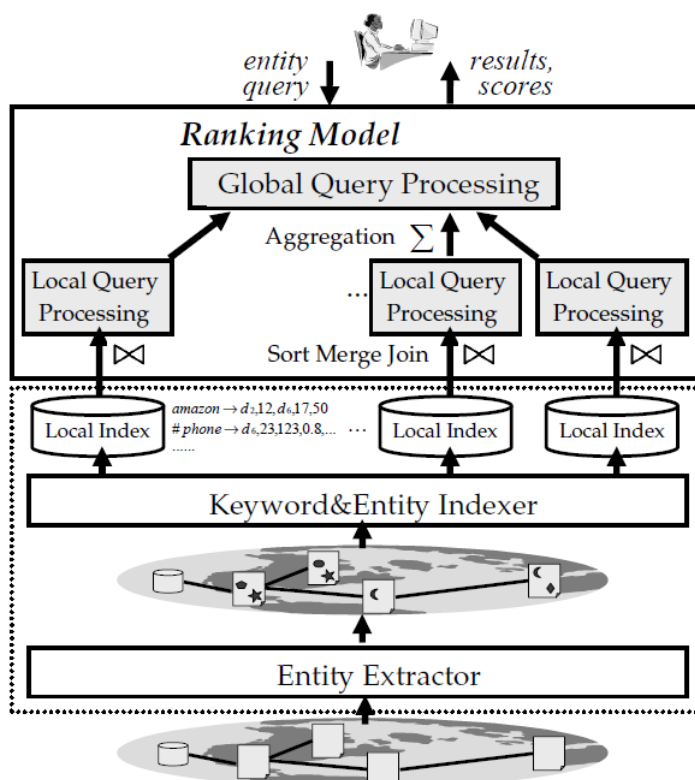


Figure 1 Cheng et al.: Entity search system architecture (taken from [Che07a])

Entity (and keyword) indexing: For the purpose of entity retrieval, Cheng et al. indexed entities as well as keywords. Specifically, the indexer in their system builds an inverted index of entities in such a way that, given an entity type, the index will return a list containing all the information concerning the entity instances extracted for the specified type (i.e., the three features recorded for each occurrence of each entity instance). Such information is stored in a list ordered by

document ID, similarly as in the keyword inverted index. Online query processing is then done by loading and processing the two inverted indices.

Cheng et al. evaluated the accuracy of entity retrieval using their system with a large-sized corpus (consisting of 48974 websites and 93 million pages) but with respect to relatively easy cases of entity extraction/retrieval, i.e., phone numbers and email addresses. (Since the evaluation mainly involved comparing the performance of different ranking algorithms, the details are not described here.)

While Cheng et al.'s project shares motivations and goals similar to those of the research involving the PanAnthropon project, there are distinct differences between the approach taken in the former and that in the latter.

First of all, it must be noted that Cheng et al.'s approach is not fully entity-centric or semantics-based, in that Cheng et al.'s entity view of the Web still retains the viewpoint of the word-based, document-oriented framework of traditional information retrieval. According to their view, the Web is a collection of entities *over a collection of documents*. As such, entity search is conceived of as *search by context*, where "context" does not mean the semantic context that conditions entities, but rather the textual context that is defined by the occurrences of specified *keywords*. Information is extracted/recorded, not for an entity instance, but for each textual occurrence thereof. The main type of information concerning an entity instance occurrence has to do with a document-centric feature, i.e., the position in a given document of the word corresponding to the entity instance. The inverted index containing such information is then organized by document IDs. Lastly, query–result matching/ranking is based on computing how well a result entity tuple conforms to the expected textual pattern across every document in the corpus.

Another way to point to the non-entity-centric, non-semantics-based nature of Cheng et al.'s approach is by noting that they lack any conceptual scheme (i.e., an ontology or taxonomy) to organize entity types and to define relevant attributes accordingly. Their rule-driven entity extractor extracts entities by using textual patterns to be matched by entities of a given type (e.g., ddd-ddd-dddd for #phone type entities, where d stands for a numeric digit). It is not clear how their dictionary-driven extractor works, but the term "dictionary" implies a word-based method. In Cheng et al.'s entity search system, an entity type is specified simply by prefixing a word with #. (Since no description is provided as to how the search interface is designed, here it is assumed that the system uses a free-text-based input method for entering a query tuple.) As such, what is conceptually one and the same entity type may be entered in a variety of ways (e.g., #phone, #phone_number, #telephone number), not to mention the fact that there can be any number of ad-hoc entity types. It is not clear if and how Cheng et al.'s system can deal with an unrestricted variety of entity types, given the fact that extraction and indexing of entities are done offline. The query tuple examples provided by Cheng et al., which are more or less like familiar Google queries, evidently show the non-semantic nature of their approach, in that no knowledge of the semantic relations among the tuple components is implied (e.g., that Amazon is a company, that a company consists of departments, that customer service is a kind of department, that a department has phone number(s), etc.).

In summary, entity search/retrieval in Cheng et al.'s approach still operates in a *textual space* within the framework of traditional document-centric information retrieval. In contrast, the type of entity retrieval that is aimed at in the PanAnthropon project operates in a *semantic space*, based on a conceptual scheme under which information concerning the entities is organized. Both the offline tasks of entity extraction and indexing and the online tasks of entity search and retrieval, in this research, are semantics-based rather than text-based, although entity extraction

requires some string pattern matching of cue words/phrases (not entity names themselves) when processing unstructured text. The commonalities between the two approaches are then only that in both approaches entity extraction/indexing is done offline and that entity search/retrieval is done via explicit specification of entity type(s) against the information stored.

The problem of ranking (related) entities of various heterogeneous types (some generic, others specific) identified from the documents returned by a search engine in response to ad-hoc keyword queries has been investigated by Zaragoza et al. [Zar07].

Zaragoza et al.'s study consisted of three phases: They first used a statistical entity extractor to extract entities and identify their corresponding types from the English version of Wikipedia. They then asked users to issue queries to a baseline entity ranking system and to manually evaluate the query results. Lastly, they compared the performance of different entity ranking algorithms on those same queries.

In order to extract entities from Wikipedia, Zaragoza et al. first trained a statistical entity extractor on the BBN Pronoun Coreference and Entity Type Corpus [Wei05], which includes annotation of 12 named entity types (Person, Facility, Organization, GPE, Location, Nationality, Product, Event, Work of Art, Law, Language, and Contact-Info), 9 nominal entity types (Person, Facility, Organization, GPE, Product, Plant, Animal, Substance, Disease, and Game), and 7 numeric types (Date, Time, Percent, Money, Quantity, Ordinal, and Cardinal). Zaragoza et al. then applied the extractor on an XML-ized Wikipedia corpus [Den06] that contains 625,405 Wikipedia entries, thereby identifying 28 million occurrences of 5.5 million unique entities. (The resulting semantically-annotated Wikipedia corpus is described in [Ats08].) Lastly, they created a retrieval index that contains the text and the identified entities and types.

In the second phase of their study, Zaragoza et al. used 10 users to issues queries and evaluate the results. Each user was asked to choose query topics that were familiar to the user and were covered in Wikipedia. The queries were then run by the system using a standard passage retrieval algorithm that retrieved the 500 most relevant passages and collected all the entities that appeared in those passages. The collected entities were then ranked by using a baseline entity ranking algorithm and were given to the user to evaluate by assigning one of five judgment labels: Most Important, Important, Related, Unrelated, and Don't know. Zaragoza et al. obtained a total of 50 judged queries through the procedure as described above.

Some of the queries and corresponding judgments of query results are shown in Table 1. With these examples, Zaragoza et al. stress the difficulty and subjectivity involved in the evaluation task. In this regard, they admit that their evaluation task design in this explorative study may have been quite naïve.

**Table 1 Zaragoza et al.: Sample queries and entity relevance judgments
(taken from [Zar07])**

Query	“Yahoo! Search Engine”
Most Important Entities	Yahoo, Google, MSN, Inktomi, Yahoo.com.
Important Entities	Web, crawler, 2004, AltaVista, 2002, Amazon.com, Jeeves, TrustRank, WebCrawler, Search Engine Placement, more than 20 billion Web, eBay, Worl Wide Web, BT OpenWorld, between 1997 and 1999, Stanford University and Yahoo, AOL, Kelkoo, Konfabulator, AlltheWeb, Excite.
Related Entities	users, Firefox, Teoma, LookSmart, Widget, companies, company, Dogpile, user, Searchen Networks, MetaCrawler, Fitzmas, Hotbot, ...
Query	“Budapest”
Most Important Entities	Budapest, Hungary, Hungarian, city, Greater Budapest, capital, Danube, <i>Budapesti Közgazdaságtudományi és Államigazgatási Egyetem</i> , M3 Line, Pest county.
Important Entities	University of Budapest, Austria, town, Budapest Metro, Soviet, 1956, Ferenc Joachim, Karl Marx University of Economic Sciences, Budapest University of Economic Sciences, Eötvös Loránd University of Budapest, ...
Related Entities	Paris, Vienna, German, Prague, London, Munich, Collegium Budapest, government, Jewish, Nazi, 1950, Debrecen, 1977, M3, center, Tokyo, World War II, New York, Zagreb, Leipzig, population, residences, state, cementery, Serbian, ...
Query	“Tutankhamun curse”
Most Important Entities	Tutankhamun, Carnarvon, mummies, Boy Pharaoh, The Curse, archaeologist, Howard Carter, 1922.
Important Entities	Pharaohs, King Tutankhamun.
Related Entities	Valley, KV62, Curse of Tutankhamun, Curse, King, Mummy’s Curse, ...

In the final phase of their study, Zaragoza et al. evaluated several different models of two different ranking algorithms — one exploiting entity containment graphs and the other using a Web search engine to compute entity relevance. (Again, the details involving the different ranking algorithms and the evaluation of their relative performance will not be addressed here.)

Zaragoza et al.'s project is both farther from and closer to the PanAnthropon project, in terms of its goal and approach, when compared with Cheng et al.'s project.

On the one hand, Zaragoza et al.'s project is farther from this research, in that, unlike the latter and unlike Cheng et al.'s project, it is not aimed at directly retrieving entities that match queries that specifically search for entities that meet certain (semantic or contextual) conditions but at identifying and ranking relevant (i.e., possibly somehow related) entities from the documents retrieved in response to ad-hoc topical queries. As the query examples in Table 1 clearly show, and as is common with many general Web search queries, such ad-hoc queries are rather open-ended in nature (by necessity or by preference) in terms of the implicit information needs. For example, when a query such as “Budapest” is entered into a search engine, the intention behind the query and the information sought thereby can be of various kinds, such as identifying the geographical location of the city, finding available options for traveling to the city, learning about the history of the city, searching for images, books, songs, films, etc. about the city, and so on. Accordingly, the relations between (the entity implied in) a query topic and the retrieved/ranked entities that are deemed relevant are also open-ended.

On the other hand, Zaragoza et al.'s project is also closer than Cheng et al.'s to this research, in that it is also based on a certain conceptual scheme to semantically categorize different types of entities (i.e., the annotation scheme from the BBN Pronoun Coreference and Entity Type

Corpus). This is in contrast to Cheng et al.'s project where such a conceptual scheme is absent, and where, accordingly, the types of entities addressed — if they can be addressed at all — remain open-ended and unorganized by default.

Nevertheless, both Zaragoza et al.'s project and Cheng et al.'s project operate within the document-centric framework of traditional information retrieval. The PanAnthropon project stands out apart from both of them, in that its overarching entity-centric, semantics-based approach encompasses the entire process from entity extraction and indexing to entity search and retrieval.

It may be noted that the task of retrieving related entities addressed by Zaragoza et al. is implicitly incorporated in the PanAnthropon interface, in a non-ambiguous, semantics-based way, in that the query results explicitly display the relations between a given entity and other entities. The interface also provides functions of retrieving direct and indirect relations between two specified entities, as will be described later.

The task of retrieving and ranking semantic-type-specified entities that answer a given query using an XML-ized Wikipedia corpus has been taken up by the INitiative for Evaluation of XML Retrieval (INEX) (<http://inex.is.informatik.uni-duisburg.de/2007/>), which in 2007 started the XML Entity Ranking (XER) Track [deV08] [Dem09].

The XER Track involves two main tasks, namely, Entity Ranking (ER) and List Completion (LC). These tasks concern information needs represented as tuples in the form of `<query, category, entity>`. The `query` component consists of free-text title and description of the topic. The `category` component specifies the type(s) of entities to be retrieved, represented by

Wikipedia categories. Finally, the `entity` component provides example instances of the specified entity type(s). The input for the ER task consists of the `query` and `category` portions of the triple, whereas the input for the LC task consists of the `query` and `entity` components. In both tasks, the entity retrieval system should return the target entities, represented by corresponding Wikipedia pages.

The Entity Relation Search (ERS) task, introduced in 2008 as a pilot task for the XER Track, builds upon the ER task, and consists of the ER phase and the ERS proper phase. The aim of this task is to find entities that are in a specified relation with respect to the entities retrieved as the result of the ER task. (For example, having found the museums in the Netherlands that exhibit Van Gogh's art works, one may wish to find the cities in which those museums are located.) Information needs for the ERS task are represented as tuples of the format `<query, category, entity, relation-query, target-category, target-entity>`, where the first three components are defined as in the ER and LC tasks. The `relation-query` component consists of free-text title and description/narrative of the relation between the entities returned in the ER phase and the target entities. The `target-category` component specifies the type(s) of the target entities to be retrieved. The `target-entity` provides example instances of the target entity type(s).

The training and testing sets for the XER Track are constructed from the topics submitted by the track participants. The INEX 2008 XER Track testing set consisted of 35 topics, such as one shown in Figure 2.

```

<inex_topic topic_id="108">
  <title>State capitals of the United States of America</title>
  <description>
    I want a list of the state capitals of the United States of America
  </description>
  <narrative>
    Each result should be an article about a capital city of a state of the United States of America.
  </narrative>
  <categories>
    <category id="1701">u.s. state capitals</category>
    <category id="10481">capitals</category>
    <category id="89169">capital cities</category>
  </categories>
  <entities>
    <entity id="17653">Lincoln, Nebraska</entity>
    <entity id="6503">Concord, New Hampshire</entity>
    <entity id="57700">Tallahassee, Florida</entity>
    <entity id="57864">Cheyenne, Wyoming</entity>
    <entity id="44186">Providence, Rhode Island</entity>
  </entities>
  <entity-relation>
    <relation-title>capital of</relation-title>
    <relation-description>
      I want the states of which these cities are capitals
    </relation-description>
    <relation-narrative>
      Each result should be an article about a United States state.
    </relation-narrative>
    <target-categories>
      <category id="471">states of the united states</category>
    </target-categories>
    <entity-pairs>
      <entity-pair>
        <main-entity id="57864">Cheyenne, Wyoming</main-entity>
        <target-entity id="33611">Wyoming</target-entity>
      </entity-pair>
      <entity-pair>
        <main-entity id="6503">Concord, New Hampshire</main-entity>
        <target-entity id="21134">New Hampshire</target-entity>
      </entity-pair>
      <entity-pair>
        <main-entity id="17653">Lincoln, Nebraska</main-entity>
        <target-entity id="21647">Nebraska</target-entity>
      </entity-pair>
      <entity-pair>
        <main-entity id="57700">Tallahassee, Florida</main-entity>
        <target-entity id="10829">Florida</target-entity>
      </entity-pair>
      <entity-pair>
        <main-entity id="44186">Providence, Rhode Island</main-entity>
        <target-entity id="25410">Rhode Island</target-entity>
      </entity-pair>
    </entity-pairs>
  </entity-relation>
</inex_topic>

```

Figure 2 INEX 2008 XER Track: Sample query topic in testing set

Table 2 (continued on the next page) shows the IDs, titles, and categories corresponding to the 35 query topics that comprised the testing set for the INEX 2008 XER Track. (Note that the category names are uniformly given in lowercase letters, presumably to clearly distinguish them from entity names. Similarly, class names and attribute names are represented in lowercase letters in this research.)

Table 2 INEX 2008 XER Track: Query topics and entity types in testing set

ID	Title	Categories
104	Harry Potter Quidditch Gryffindor character	harry porter harry porter characters
106	Noble English person from the Hundred Years' War	peerage of England hundred years' war
108	State capitals of the United States of America	u.s. state capitals capitals capital cities
109	National capitals situated on islands	capitals
110	Nobel Prize in Literature winners who were also poets	nobel prize in literature winners
112	Guitarists with mass-produced signature guitar models	guitarists
113	Formula 1 drivers that won the Monaco Grand Prix	racecar drivers formula one drivers
114	Formula one races in Europe	formula one grands prix
115	Formula One World Constructors' Champions	formula one constructors
116	Italian nobel prize winners	nobel laureates
117	Musicians who appeared in the Blues Brothers movies	musicians
118	French car models in 1960's	automobile manufacturers french automobile manufacturers
119	Swiss cantons where they speak German	geography of switzerland cantons of switzerland
121	US presidents since 1960	presidents of the united states u.s. democratic party presidential nominees u.s. republican party presidential nominees
122	Movies with eight or more Academy Awards	best picture oscar british films american films
123	FIFA world cup national team winners since 1974	football in brazil european national football teams football in argentina
124	Novels that won the Booker Prize	novels
125	countries which won the FIFA world cup	countries
126	toy train manufacturers that are still in business	toy train manufacturers

Table 2 (continued)

127	german female politicians	politicians women
128	Bond girls	film actors bond girls
129	Science fiction book written in the 1980	science fiction novels science fiction books
130	Star Trek Captains	star trek: the next generation characters star trek: voyager characters star trek: deep space nine characters
132	Living nordic classical composers	21st century classical composers finnish composers living classical composers
133	EU countries	country
134	record-breaking sprinters in male 100-meter sprints	sprinters
135	professional baseball team in Japan	japanese baseball teams
136	Japanese players in Major League Baseball	baseball players
138	National Parks East Coast Canada US	national parks national parks of the united states national parks of canada
139	Films directed by Akira Kurosawa	japanese films
140	Airports in Germany	airports in germany
141	Universities in Catalunya	catalan universities
143	Hanseatic league in Germany in the Netherlands Circle	cities cities in germany
144	chess world champions	chess grandmasters world chess champions
147	Chemical elements that are named after people	eponyms

The major difference between most of the approaches used by the participants in the INEX XER Track (e.g., [Dem08] [Jäm08] [Jia09] [Kap09] [Mur08] [Shi08] [Tsi08] [Ver08] [Wee09] [Zhu08]) and those of Cheng et al., Zaragoza et al., and the PanAnthropon project, consists in the fact that the former do not involve the offline processing of entity extraction/indexing, so that entity retrieval/ranking is performed directly against the corpus itself, whereas, in the latter, entity extraction/indexing is done offline by extracting entity instances and by recording information on those entities prior to the online processing of entity retrieval (and ranking), so that entity retrieval/ranking is performed against the stored information concerning the entities.

The INEX XER Track approaches, however, are also somewhat similar to Zaragoza et al.'s approach, as far as the process of entity search/retrieval/ranking is concerned, in that both approaches first retrieve a ranked list of documents/passages that are potentially relevant to the given query topic by using a more-or-less standard document/passage retrieval engine/algorithm based on keyword matching, then identify relevant/matching entities from the set of documents retrieved/ranked, and finally rank the retrieved entities by some scores to measure the degree of relevance/matching. As such, entity retrieval/ranking in both approaches is a procedure added onto the standard document retrieval process, not its replacement. The differences between the two approaches consist in the fact that, in the case of the INEX XER Track, the types of entities to be retrieved (in the ER task), the corresponding entity instances (in the LC task), or the relations between the initially retrieved entities and the target entities (in the ERS task) are explicitly specified for the purpose of entity retrieval/ranking, and that only the entities, and not the documents, are the real targets of queries (although the entities are represented by the corresponding Wikipedia pages).

In contrast, in this research and Cheng et al.'s, entity retrieval/ranking replaces document retrieval/ranking entirely, in that retrieving/ranking entities that match a given query is done directly over the entities extracted/stored, without the intermediate process of retrieving/ranking documents.

Given that most of the approaches used in the INEX XER Track perform entity retrieval/ranking on top of document retrieval/ranking, those approaches focus on devising the methods and scoring schemes to improve the accuracy of initial document retrieval/ranking and that of subsequent entity retrieval/ranking. Several approaches [Jia09] [Shi08] [Tsi08] [Wee09] are statistically-oriented, in that they use some probabilistic models (such as random-walk models

and generative language models) to estimate the relevance of documents/entities. Others [Mur08] [Zhu08] focus on the extraction/combination of textual content-based document features to be used in retrieving/ranking documents/entities. Still others [Dem08] [Jäm08] [Kap09] [Ver08] focus heavily on the structural features of Wikipedia pages, i.e., the links and categories therein, for query expansion/refinement and relevance propagation/calculation.

One system that stands out among the INEX XER Track systems, which is more relevant to this research, is one by Craswell et al. [Cra09]. Even though two official runs they submitted to the INEX 2008 XER Track evaluations used approaches that are more or less similar to those of the other systems, Craswell et al. describe another explorative approach based on structured indexing of entities, which is quite relevant to this research.

By structured indexing of entities, Craswell et al. mean generating a set of `<attribute, value>` pairs to represent the information concerning the entities. What Craswell et al. attempted in their third approach is to generate such a structured entity index from unstructured (portions of) Wikipedia pages. The process of generating the entity index consisted of two steps: entity reference resolution and attribute extraction.

Entity reference resolution: In order to build the structured entity representation, Craswell et al. first had to detect the occurrences of the references to the entities from the Wikipedia pages. They performed the entity reference resolution task in two different ways, depending on whether a given page is a page representing the given entity or a page linking to the page representing the entity (e.g., a list-of page). In the first case, Craswell et al. consider all occurrences of the terms similar to the page title as entity references. (For example, in a page titled “Napoleon (1995 film)”, all occurrences of “Napoleon” or “Napoleon (1995 film)” are considered as entity

references.) In the second case, Craswell et al. consider the anchor text of a link as an entity occurrence.

Attribute extraction: Using the Stanford Parser (<http://nlp.stanford.edu/software/lex-parser.shtml>) and manually-created rules on the grammatical structure, Craswell et al. extracted attribute names and values from the Wikipedia pages representing entities. The rules they created cover two general sentence structure cases of the format <entity reference, attribute name, attribute value> and of the format <attribute value, attribute name, entity reference>, i.e., the case in which the entity reference is the subject of the sentence and the case in which it is the object of the sentence.

Figure 3 shows the rules Craswell et al. created using the syntax of the Tregex pattern matching tool (<http://nlp.stanford.edu/software/tregex.shtml>). (In the rules, `word` represents an entity reference, `name` represents the attribute name, and `value` represents the attribute value.)

```

S < ((NP << word) $. (VP << (VBZ|VBD|VBG|VBN|VBP = name1 . (
    VP < (TO = name2 $. (VP = value))))))
S < ( (NP << (NP = name $. (PP < (IN $. (NP << word)))))) $. (
    VP << (TO $. VP = value))
S < ('' $. (NP << word $. (''$. (VP << (
    VBZ|VBD|VBG|VBN|VBP = name $. (
    NP = value1 ?$. (/,/ $. PP = value2))))))
NP << word . (TO $. (VP < (VB = name $. NP = value)))
S < (NP << word $. (VP < (VBZ|VBD|VBG|VBN|VBP = name $. (
    PP = value1 $. PP = value2))))
word . (PP < (IN = name . NP = value))
S < (NP << word $. (VP < (VBZ|VBD|VBG|VBN|VBP = name1 $. (
    ADVP = name2 $. (VP < VBZ|VBD|VBG|VBN|VBP = name3 < (PP < (
    IN = name4 $. NP = value1 ?$. (/,/ $. PP = value2))))))))
S < ((NP << word) ?$. VBZ|VBD|VBG|VBN|VBP = name1 ?$. .
    ADVP = name2 $. . (VP << (VBZ|VBD|VBG|VBN|VBP = name2 $. (
    NP = value1 ?$. (/,/ $. PP = value2))))

```

**Figure 3 Craswell et al.: Pattern matching rules for attribute extraction
(taken from [Cra09])**

In addition to the grammar rules shown in Figure 3, Craswell et al. also used some ad-hoc rules for the attribute extraction task. By using both types of rules, Craswell et al. extracted attributes for 368,788 entities. Table 3 shows what Craswell et al. consider as meaningful, representative examples of the attributes extracted.

Table 3 Craswell et al.: Examples of entity attributes extracted (taken from [Cra09])

Entity	Attribute	Value
Albert Einstein	was	a Jewish German - born theoretical physicist of profound genius , who is widely regarded as the greatest scientist of the 20th century
Albert Einstein	described	the “predatory phase of human development”
Lausanne	is	a city in the French-speaking part of Switzerland, situated on the shores of Lake Geneva (French: Lac Lman), and facing vian-les-Bains (France)
Lausanne	located	some 60 km northeast of Geneva
Lausanne	follows	“La Nuit de Muses” (Museum’s night) in the fall season
Lausanne	boasts	a dramatic panorama over the lake
Lausanne	is	the birthplace of: Umberto Agnelli, Anthony Bloom, Franois-Louis David, Bocion Johann, Ludwig Burckhardt, Benjamin Constant, Aloise Corbaz, Charles Dutoit, Egon von Furstenberg, Eugne Grasset, Bertrand Piccard, Charles Ferdinand, Ramuz Thophile, Steinlen Elizabeth, Thompson (Lady Butler), Bernard Tschumi, Flix Vallotton
Lausanne	has	some alternative culture
Lausanne	provide	a diverse and rich musical life
Lausanne	is	E 37 N 46 10 56 31 6
Lausanne	going to	become the first city in Switzerland to have a real metro system, with the m2 Line which will open in 2008

For the ERS task, Craswell et al. also indexed the relations between entities in a structured fashion. They consider two entities as being related if the references to the two entities co-occur in a sentence. For each pair of two co-occurring entity references, Craswell et al. indexed the

sentences in which they occur. Example entries of the entity relation index they created are shown in Table 4.

Table 4 Craswell et al.: Examples of entity relations extracted (taken from [Cra09])

Entity	Predicate	Entity
flint river	the flint river with an area of 568 square miles is a tributary to the	tennessee river
flint river	much of the 342 sq	watershed
albert einstein	he proposed the and also made major contributions to the development of and the theory provided the foundation for the study of and gave scientists the tools for understanding many features of the universe that were discovered well after einstein death	cosmological models
albert einstein	he proposed the and also made major contributions to the development of and	theory of relativity

Although Craswell et al.'s approach as described above is similar to that of the PanAnthropon project, insofar as the use of structured representation and indexing of entities is concerned, the methods of extracting the information to be used for constructing such structured indices as well as the formats and contents of the indices thus created differ between the two.

The methods Craswell et al. used to extract attribute names/values and entity relations are comparable to those used for open information extraction on the general Web [Ban07], in that both approaches extract (the equivalents of) <attribute, value> pairs, <entity, attribute, value> triples, or <entity, relation, entity> triples from unstructured text by using pattern matching (although there are differences in the details involved). Both approaches are open-ended in terms of the number and kinds of attributes/relations to be extracted, as a whole and per each entity.

While such approaches, which commonly generate noisy data with low precision (and low utility), may be inevitable for the purpose of open-domain information extraction from the unstructured Web pages that comprise the bulk of the Web, such is not the case when it comes to extracting information from Wikipedia, the structural features of which can be readily used to represent semantic information. Indeed, the “open information extraction” (OIE) approach in [Ban07] is (motivated by the desire to pursue an approach that is) orthogonal to knowledge-based information extraction approaches such as those often used on Wikipedia. It is thus peculiar that Craswell et al. chose to use the unstructured text portions of Wikipedia pages to build structured entity/relation indices.

The examples in Table 3 and Table 4 clearly show the low quality level of information extracted by using an “open” (i.e., knowledge-blind) approach. The attributes shown are at best trivial or meaningless and at worst misleading. (For example, `is` alone is considered as an attribute, whereas `is_birthplace_of` should have been considered as such.) The attribute values and entity-relation predicates mostly consist of rather long sentence fragments, which are not useful beyond their particular occurrences.

Craswell et al. do mention that, in the future, they will investigate leveraging the information presented in semi-structured formats (e.g., infoboxes) in Wikipedia. They envision that, by doing so, they will be able to perform entity retrieval/ranking by using an inverted index built on top of the attribute values for all identified entities. What is left as a future work by Craswell et al., namely, extracting semi-structured information from Wikipedia and enabling entity retrieval by structured queries, each of which represented as a set of `<attribute, value>` pairs or `<entity, attribute, value>` triples, based on the structured representation of information

concerning the entities, has already been attempted by, e.g., Auer et al. [Auer07a] [Auer07b] and Suchanek et al. [Suc07] [Suc08] (discussed in Sections 4.2 and 4.3).

The approach used in this research is distinct from both Craswell et al.'s (current) approach and those of Auer et al. and Suchanek et al., in that, while it mainly takes advantage of the semi-structured infoboxes and categories in Wikipedia pages to extract semantic information, as in the latter, it also uses the unstructured portions of Wikipedia pages for the same purpose, as in the former, but in a knowledge-aware manner with a predetermined set of attributes, thereby extracting/deriving uniformly structured/formatted entity-centric facts of high quality to be used for direct entity retrieval.

To sum up the differences between the approach of the INEX XER Track in general and that of the PanAnthropon project: Unlike in the INEX XER Track, where entity retrieval/ranking is done as an added step after query-dependent, keyword-based document retrieval/ranking is performed on an XML-ized Wikipedia corpus, this research aims to enable entity retrieval without the intermediate step of document retrieval, by first building a repository of semantic information on the entities by extracting/deriving `<attribute, value>` pairs from a subset of native HTML Wikipedia pages. Also, unlike in the XER Track, the entity search interface created from this research uses a form-based method for query input. This research focuses on the entity retrieval capability that is comparable to the main Entity Ranking task in the XER Track (without the ranking part). However, a function analogous to, but even more general than, the Entity Relation Search task, is also implicitly provided by the way the query results are presented.

4.2 Information Extraction from Wikipedia

Wikipedia has in recent years become a topic of intense interest among computer and information science researchers involved with various research areas such as natural language processing, information extraction, information retrieval, knowledge engineering, Semantic Web, social network analysis, etc. While there have been some approaches geared toward enhancing Wikipedia by introducing/augmenting semantic features, most of the approaches have focused on exploiting Wikipedia as a lexical, topical, and semantic knowledge resource for various tasks and applications.

For example, Wikipedia has been used or examined in the context of a variety of areas, topics, and applications such as:

- Characteristics of Wikipedia as a knowledge resource (e.g., [Nak08c] [Zes07b])
- Class–instance differentiation (e.g., [Zir08])
- Community content creation (e.g., [Hof09] [Wel08])
- Community detection (e.g., [Liz09])
- Concept mapping (e.g., [Med08a] [Pon09] [Rei08])
- Coreference resolution (e.g., [Pon06])
- Document classification (for Wikipedia) (e.g., [Gan09])
- Document classification (using Wikipedia) (e.g., [Jan07] [WaP07] [Wea06])
- Document summarization (e.g., [Nas08a] [Sau09])
- Domain concept model construction (e.g., [Tho08])
- Gazetteer generation (e.g., [Tor06] [Zha09])
- Graph analysis (e.g., [Zes07a])
- Keyword extraction (e.g., [Gri09a] [Gri09b])
- Lexical reference (e.g., [Shn09])

- Link analysis (e.g., [Ada05])
- Link generation (e.g., [Mil08b])
- Named entity classification (e.g., [Bus07])
- Named entity disambiguation (e.g., [Bun06] [Cuc07] [Rah08])
- Named entity recognition (e.g., [Bal09] [Kaz07] [Not08])
- Network analysis (e.g., [Bel05] [Zla06])
- Ontology/taxonomy construction (e.g., [Cui08] [Ped08] [Pic07] [Pon07a])
- Ontology/taxonomy evaluation (e.g., [Yu07])
- Ontology/taxonomy extension (e.g., [Sar09] [Suc09])
- Ontology/taxonomy mapping (e.g., [Pon09])
- Ontology/taxonomy refinement (e.g., [Wu08])
- Question answering (e.g., [Ahn04] [Bus06] [Kai08])
- Semantic annotation (e.g., [Sch08])
- Semantic authoring (e.g., [Fu07])
- Semantic relatedness computation (e.g., [Mil07] [Mil08a] [Pon07b] [Str06] [Yeh09])
- Semantic relation extraction (from unstructured text) (e.g., [Blo07] [Cul06] [Ift08] [Nak08b] [Ngu07] [Rui06] [Rui07] [Suc06a] [Suc06b] [WaG07] [Yan09])
- Semantification of Wikipedia (e.g., [Kr05] [Kr07] [Völ06] [Vra06] [Wu07])
- Sentence compression (e.g., [Yam08])
- Tag classification (e.g., [Ove09])
- Term–concept mapping (e.g., [Rui05])
- Term–concept network construction (e.g., [Gre06])
- Thesaurus construction (e.g., [Mil06] [Mil07a] [Nak07a] [Nak07b])
- Thesaurus extension (e.g., [Med08b])
- Timeline generation for named entities (e.g., [Bho07])

- Topic identification/indexing (e.g., [Cou09a] [Cou09b] [Med08c] [Sch06] [Sye07])
- Topic map construction (e.g., [Yang07])
- Word sense disambiguation (e.g., [Mih07])

As stated in Chapter 3, this dissertation research proposes to exploit Wikipedia as a knowledge resource for information extraction and aims to demonstrate an improved information retrieval capability. The review of related works in this section will therefore focus on information extraction from Wikipedia; the following section will cover the information retrieval part.

Information extraction from Wikipedia (or in general) involves three major aspects of consideration: *what*, *where*, and *how*, i.e., the type of information that is the target of information extraction, the source/ location from which such information is to be extracted, and the method via which information extraction is to be performed. In general, the *where* depends on the *what*, and the *how* depends on the *where*.

The PanAnthropon project is mainly concerned with extracting/deriving semantic/ontological information (explicitly or implicitly) represented in the form of `<attribute, value>` pairs or `<entity, attribute, value>` triples, where `value` corresponds to a literal, another non-class entity, a class, or a Wikipedia category.

What renders Wikipedia a particularly useful (albeit not the only possible) resource for the purpose of this research consists in the fact that Wikipedia contains not only textual content but structural features (e.g., infoboxes, categories, etc.) which facilitate extraction of semantic information *with respect to* the entity represented by the given Wikipedia page. Moreover, the kinds of attributes that appear in infoboxes are more or less standardized according to the domain

and entity type, so that it is possible to uniformly extract *homogeneous* types of facts for all entities of a given type within the same domain.

As such, Wikipedia lends itself readily to structural mining, i.e., extracting information by exploiting the structural features of the information source, which is contrasted with content mining, which mainly uses the textual content, to employ the distinction made in [Zes07b].

The types of works that are most relevant to the PanAnthropon project, as far as information extraction is concerned, are therefore those that are mainly concerned with structural mining of Wikipedia for the purpose of extracting entity-centric facts that can be stored/retrieved using a structured semantic representation. In this sense, Auer et al.'s project concerning the DBpedia knowledge base [Auer07a] [Auer07b] and Suchanek et al.'s concerning the YAGO (Yet Another Great Ontology) knowledge base [Suc07] [Suc08], both of which are Semantic Web projects, are closely related to this research.

Suchanek et al.'s YAGO (<http://www.mpi-inf.mpg.de/yago-naga/yago/index.html>) project is concerned with building a large-scale, extensible, ontological knowledge base consisting of entities, (both taxonomic and non-taxonomic) relations, and facts, based on the information extracted from (the XML dumps of) the English version of Wikipedia and united with the information in the lexical semantic knowledge base WordNet (<http://wordnet.princeton.edu/>) [Fel98], by using a combination of rule-based and heuristics-based methods. By using both Wikipedia and WordNet, Suchanek et al. intended to benefit from the respective advantages of the two resources, i.e., the vast number of individuals (i.e., common entities which are neither facts, nor relations, nor classes) in Wikipedia, on the one hand, and the clean, well-defined taxonomy of concepts in WordNet, on the other.

YAGO is built upon a data model within the knowledge representation framework of description logics [Baa03]. The YAGO data model extends the RDFS (<http://www.w3.org/TR/rdf-schema/>) formalism, and is intended to be able to express entities, facts, relations between facts, and properties of relations.

In the YAGO data model, as in OWL (<http://www.w3.org/TR/owl-features/>) and RDFS, all objects are represented as entities. Two entities can stand in a relation (e.g., `<AlbertEinstein hasWonPrize NobelPrize>`). Literals are represented as entities (e.g., `<AlbertEinstein bornInYear 1879>`). Furthermore, words are also considered as entities (e.g., `<"Einstein" means AlbertEinstein>`). Each and every entity is an instance of at least one class (e.g., `<AlbertEinstein type physicist>`).

Classes are also considered as entities in the YAGO model. As such, each class is an instance of a class, i.e., the class `class`. Classes constitute a taxonomic hierarchy (i.e., a subsumption hierarchy), explicitly represented by the `subClassOf` relation (e.g., `<physicist subClassOf scientist>`). As literals are considered as proper entities in the YAGO data model, there is also a hierarchy of classes corresponding to different types of literals, all subsumed under the class `Literal`, as shown in Figure 4.

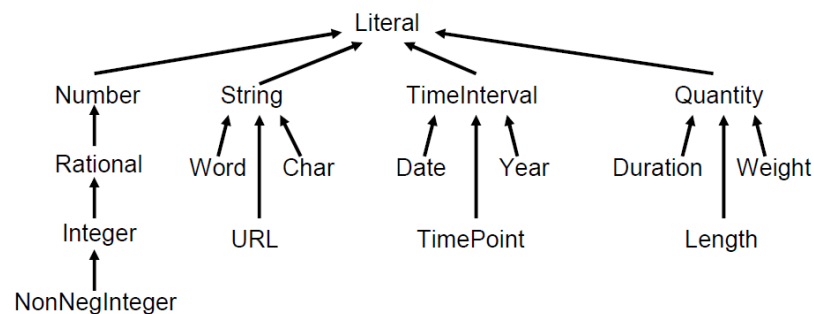


Figure 4 YAGO: Literal classes in data model (taken from [Suc08])

The YAGO model considers relations as entities, so that properties of relations (e.g., transitivity or subsumption) can also be represented (e.g., `<subClassOf type transitiveRelation>`).

In the YAGO model, a fact is represented by the triple of the form `<entity, relation, entity>`. Not only is each fact given a unique fact ID, each fact itself is also considered as an entity, so that a fact about a given fact can be represented by referring to the fact ID of the latter (e.g., `<#1 foundIn http://en.wikipedia.org/Einstein>`).

Suchanek et al.'s solution to the problem of representing arbitrary n -ary relations (where $n > 2$) is based on the assumption that, for each such relation, a primary pair of its entity arguments can be identified. The problem is thus tackled by first representing the relation between the arguments in the primary pair as a binary fact and then representing each of the other arguments with an additional binary fact that represents the relation between the primary fact and the argument at issue. For example, the fact that Albert Einstein won the Nobel Prize in 1921 can be represented by two triples, i.e., the primary fact of `<AlbertEinstein hasWonPrize NobelPrize>`, which is given the ID #1, and an auxiliary fact of `<#1 time 1921>`. In this way, an n -ary relation will be represented with $n-1$ fact triples.

Figure 5 shows axiomatic rules in the YAGO data model, which represent the `type`, `domain`, `range`, and `subClassOf` relations involving some of the classes and relations. (For example, `subClassOf` belongs to the class of `acyclicTransitiveRelation` and takes `class` both as its domain and as its range.)

```

∅ ↦ (domain, DOMAIN, relation)
∅ ↦ (domain, RANGE, class)
∅ ↦ (range, DOMAIN, relation)
∅ ↦ (range, RANGE, class)
∅ ↦ (subClassOf, TYPE, acyclicTransitiveRelation)
∅ ↦ (subClassOf, DOMAIN, class)
∅ ↦ (subClassOf, RANGE, class)
∅ ↦ (type, RANGE, class)
∅ ↦ (subRelationOf, TYPE, acyclicTransitiveRelation)
∅ ↦ (subRelationOf, DOMAIN, relation)
∅ ↦ (subRelationOf, RANGE, relation)
∅ ↦ (boolean, SUBCLASSOF, literal)
∅ ↦ (number, SUBCLASSOF, literal)
∅ ↦ (rationalNumber, SUBCLASSOF, number)
∅ ↦ (integer, SUBCLASSOF, rationalNumber)
∅ ↦ (timeInterval, SUBCLASSOF, literal)
∅ ↦ (dateTime, SUBCLASSOF, timeInterval)
∅ ↦ (date, SUBCLASSOF, timeInterval)
∅ ↦ (string, SUBCLASSOF, literal)
∅ ↦ (character, SUBCLASSOF, string)
∅ ↦ (word, SUBCLASSOF, string)
∅ ↦ (URL, SUBCLASSOF, string)

```

Figure 5 YAGO: Axiomatic rules in data model (taken from [Suc07])

A characteristic feature of Suchanek et al.’s approach to semantic/ontological knowledge extraction from Wikipedia consists in the fact that it does not only use infoboxes to extract `<attribute, value>` pairs for the individual represented by the page but it also heavily uses categories in order to extract classes and facts involving taxonomic and non-taxonomic relations. In this regard, Suchanek et al. distinguish among different types of categories found in Wikipedia: conceptual, relational, thematic, and administrative.

Conceptual categories refer to those that identify a class for the entity corresponding to the given page (e.g., “Naturalized citizens of the United States”). Relational categories refer to those that contain certain relational information with respect to the given entity individual (e.g., “1879 births”). Thematic categories refer to those that simply indicate thematic vicinity (e.g., “Physics”). Finally, administrative categories refer to those that only serve administrative

purposes for the management of Wikipedia. Suchanek et al. used different types of categories in order to extract facts involving different kinds of relations.

Extraction of type relation facts: Only conceptual categories are considered as candidates for identifying a class for the individual. In order to identify conceptual categories, Suchanek et al. excluded relational and administrative categories by manual inspection. In order to distinguish between conceptual categories and thematic categories and to extract class names from conceptual categories, Suchanek et al. used shallow linguistic parsing and stemming of category names (using the Noun Group Parser and Pling-Stemmer in [Suc06a]). The heuristics they used is based on the observation that, if the head of a category name is a plural word, the category is most likely to be a conceptual category. (For example, in the category name “Naturalized citizens of the United States”, “Naturalized” is a pre-modifier, “citizens” is the head, and “of the United States” is a post-modifier. The identified head, “citizens”, is then stemmed to remove its plural ending. The class membership information for the individual can then be represented as, e.g., `<AlbertEinstein type naturalizedCitizenOfTheUnitedStates>.`)

Extraction of subclassOf relation facts: Considering that the category structure in Wikipedia, albeit hierarchically organized, does not form an ontologically well-defined taxonomy but rather only reflects the thematic structure of Wikipedia, Suchanek et al. took only the leaf categories from Wikipedia and used WordNet instead to establish the class hierarchy. Specifically, they used the hyponymy relation in WordNet to derive the subclassOf hierarchy, following the rule that a class is a subclass of another, if the first class name (or its synonym) amounts to a hyponym of the second one. (The algorithm they then used to connect a subclass from Wikipedia to a super-class in WordNet is described in [Suc07] [Suc08].)

Extraction of means relation facts: Suchanek et al. used both WordNet and Wikipedia for this purpose. When using WordNet, Suchanek et al. first created a class for each synset (synonym set) and then established the `means` relation between each word in a synset and the corresponding class (e.g., `<"metropolis" means city>`). When using Wikipedia, Suchanek et al. exploited its redirect pages to extract alternative names for individuals (e.g., `<"Einstein, Albert" means AlbertEinstein>`). They established `givenNameOf` and `familyNameOf` relations as subrelations of the `means` relation when a given individual is a person, by using the Name Parser from [Suc06a] (e.g., `<"Einstein" familyNameOf AlbertEinstein>`).

Extraction of facts involving other relations: Suchanek et al. used heuristics to extract non-taxonomic, non-definitional relations, e.g., `bornInYear`, `diedInYear`, `establishedIn`, `locatedIn`, `writtenInYear`, `politicianOf`, `hasWonPrize`, by using relational categories in Wikipedia. For example, the `bornInYear`, `diedInYear`, and `establishedIn` relations were extracted from category names ending with “births”, “deaths”, and “establishments”, respectively (e.g., “1879 births”). The `locatedIn` facts were extracted from categories such as “Countries in ...”, “Rivers of ...”, etc. Facts concerning other relations were extracted similarly.

Extraction of meta-relation facts: Besides the facts involving the ordinary relations as described above, which hold between individuals or between classes and individuals, Suchanek et al. also recorded those involving a few meta-relations: `describes`, `foundIn`, `extractedBy`, and `context`. The `describes` relation is established between the URL of a Wikipedia page and the individual represented by the page. The `foundIn` holds between a fact and the URL of the page from which the fact was extracted, while the `extractedBy` relation holds between a fact and the technique by which it was extracted. Finally, the `context` relation is established between an individual and the individuals it is linked to in Wikipedia.

YAGO was said to contain approximately 1.7 million entities and 15 million facts about those entities, as of [Suc08]. Table 5 shows the number of facts involving the various relations described above, as of the time of [Suc07]. Table 6 shows the accuracy of facts concerning those relations as judged through manual inspection of a small number of facts, also as of [Suc07]. Tables 7 and 8 show the number of entities and the number of facts for the largest relations (i.e., relations with the largest number of facts) as of [Suc08]. Table 9 shows sample facts in YAGO, as presented in [Suc07].

Table 5 YAGO: Number of facts as of 2007 (taken from [Suc07])

Relation	Domain	Range	# Facts
SUBCLASSOF	class	class	143,210
TYPE	entity	class	1,901,130
CONTEXT	entity	entity	~40,000,000
DESCRIBES	word	entity	986,628
BORNINYEAR	person	year	188,128
DIEDINYEAR	person	year	92,607
ESTABLISHEDIN	entity	year	13,619
LOCATEDIN	object	region	59,716
WRITTENINYEAR	book	year	9,670
POLITICIANOF	organization	person	3,599
HASWONPRIZE	person	prize	1,016
MEANS	word	entity	1,598,684
FAMILYNAMEOF	word	person	223,194
GIVENNAMEOF	word	person	217,132

Table 6 YAGO: Accuracy of facts as of 2007 (taken from [Suc07])

Relation	# evaluated facts	Accuracy
SUBCLASSOF	298	97.70% ± 1.59%
TYPE	343	94.54% ± 2.36%
FAMILYNAMEOF	221	97.81% ± 1.75%
GIVENNAMEOF	161	97.62% ± 2.08%
ESTABLISHEDIN	170	90.84% ± 4.28%
BORNINYEAR	170	93.14% ± 3.71%
DIEDINYEAR	147	98.72% ± 1.30%
LOCATEDIN	180	98.41% ± 1.52%
POLITICIANOF	176	92.43% ± 3.93%
WRITTENINYEAR	172	94.35% ± 3.33%
HASWONPRIZE	122	98.47% ± 1.57%

Table 7 YAGO: Number of entities as of 2008 (taken from [Suc08])

Relations	92
Classes	224,391
Individuals (without words and literals)	1,531,588
People	546,308
Locations	230,988
Institutions/companies	57,893
Movies	33,234

Table 8 YAGO: Number of facts involving largest relations as of 2008 (taken from [Suc08])

Relation	# Facts	Relation	# Facts
hasUTCOffset	12724	hasWonPrize	13645
livesIn	15185	writtenInYear	16441
originatesFrom	16876	directed	18633
hasPredecessor	19154	actedIn	22249
hasDuration	23652	bornInLocation	24400
hasImdb	24659	hasArea	26781
hasProductionLanguage	27840	produced	30519
hasPopulation	30731	isOfGenre	33898
hasSuccessor	46658	establishedOnDate	69529
hasWebsite	79779	created	83627
locatedIn	125738	diedOnDate	168037
subclassOf	211979	bornOnDate	350613
givenNameOf	464816	familyNameOf	466969
inLanguage	2389627	isCalled	2984362
type	3957223	means	4014819

Table 9 YAGO: Sample facts (taken from [Suc07])

Zidane	TYPE+SUBCLASS	football player
Zidane	TYPE	Person from Marseille
Zidane	TYPE	Legion d'honneur recipient
Zidane	BORNINYEAR	1972
"Paris"	FAMILYNAMEOF	Priscilla Paris
"Paris"	GIVENNAMEOF	Paris Hilton
"Paris"	MEANS	Paris, France
"Paris"	MEANS	Paris, Texas
Paris, France	LOCATEDIN	France
Paris, France	TYPE+SUBCLASS	capital
Paris, France	TYPE	Eurovision host city
Paris, France	ESTABLISHEDIN	-300

The YAGO model is designed to be independent of a particular data storage format. As an internal format, Suchanek et al. chose to store the extracted information using simple text files in such a way that a folder is created for each relation and each folder contains files that list all facts involving the given relation (i.e., in the form of entity arguments followed by an associated fact confidence value between 0 and 1 inclusive), as shown in Figure 6. However, they also provide programs to convert the YAGO ontology to different output formats, such XML, RDFS, and an Oracle, Postgres, or MySQL database table represented by the schema `FACTS(factID, arg1, relation, arg2, confidence)`. Figure 7 shows a snippet of the RDFS version of YAGO. Figure 8 shows a snippet of the `subClassOf` hierarchy in YAGO, also rendered in RDFS.

Sample facts from the YAGO ontology in raw text format

Folder "hasAcademicSupervisor":

400822917	Alain_Connes	Jacques_Dixmier	1.0
400822921	Albert_Einstein	Alfred_Kleiner	1.0
400822925	Andrey_Markov	Pafnuty_Chebyshev	1.0
400822929	Abram_Samoilovitch_Besicovitch	Andrey_Markov	1.0
400822933	Georgy_Voronoy	Andrey_Markov	1.0
400822937	Nikolai_Gunter	Andrey_Markov	1.0
400822957	Jacob_Tamarkin	Andrey_Markov	1.0
400822977	Augustus_De_Morgan	George_Peacock	1.0
400822981	Edward_Routh	Augustus_De_Morgan	1.0
400822985	Adolph_Wilhelm_Hermann_Kolbe	Robert_Bunsen	1.0
400822989	Peter_Griess	Adolph_Wilhelm_Hermann_Kolbe	1.0
400822993	Edward_Frankland	Adolph_Wilhelm_Hermann_Kolbe	1.0
400822997	Andrew_Wiles	John_Coates	1.0
400823001	Manjul_Bhargava	Andrew_Wiles	1.0
400823005	Kartik_Prasanna	Andrew_Wiles	1.0
400823025	Arthur_Stanley_Eddington	Horace_Lamb	1.0
400823029	Leslie_Comrie	Arthur_Stanley_Eddington	1.0
400823033	Alan_Baker	Harold_Davenport	1.0
400823037	John_Coates	Alan_Baker	1.0
400823041	Roger_Heath-Brown	Alan_Baker	1.0
400823045	David_Masser	Alan_Baker	1.0
400823049	Yuval_Flicker	Alan_Baker	1.0
400823053	Cameron_Stewart_(mathematician)	Alan_Baker	1.0
400823077	Herman_Boerhaave	Burchard_de_Volder	1.0
400823081	Gerard_Van_Swieten	Herman_Boerhaave	1.0

Figure 6 YAGO: Sample facts in raw text format (taken from the YAGO project site)

```

<!DOCTYPE rdf:RDF (View Source for full doctype...)>
- <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xml:base="http://www.mpii.de/yago/resource" xmlns:rdfs="http://www.w3.org/2000/01/rdf-
  schema#" xmlns:y="http://www.mpii.de/yago/resource/">
- <rdf:Description rdf:about="http://www.mpii.de/yago/resource/Núria_Espert">
  <y:actedIn rdf:ID="f400000001" rdf:resource="http://www.mpii.de/yago/resource/Actrius" />
</rdf:Description>
- <rdf:Description rdf:about="#f400000001">
  <y:confidence
    rdf:datatype="http://www.w3.org/2001/XMLSchema#double">0.9662078477734617</y:confidence>
</rdf:Description>
- <rdf:Description rdf:about="http://www.mpii.de/yago/resource/Mercè_Pons">
  <y:actedIn rdf:ID="f400000005" rdf:resource="http://www.mpii.de/yago/resource/Actrius" />
</rdf:Description>
- <rdf:Description rdf:about="#f400000005">
  <y:confidence
    rdf:datatype="http://www.w3.org/2001/XMLSchema#double">0.9662078477734617</y:confidence>
</rdf:Description>
- <rdf:Description rdf:about="http://www.mpii.de/yago/resource/Martin_Sheen">
  <y:actedIn rdf:ID="f400000009"
    rdf:resource="http://www.mpii.de/yago/resource/Apocalypse_Now" />
</rdf:Description>
- <rdf:Description rdf:about="#f400000009">
  <y:confidence
    rdf:datatype="http://www.w3.org/2001/XMLSchema#double">0.9662078477734617</y:confidence>
</rdf:Description>
- <rdf:Description rdf:about="http://www.mpii.de/yago/resource/Calista_Flockhart">
  <y:actedIn rdf:ID="f400000013"
    rdf:resource="http://www.mpii.de/yago/resource/Ally_McBeal" />
</rdf:Description>

```

Figure 7 YAGO: Snippet of RDFS version (taken from the YAGO project site)

```

<!DOCTYPE rdf:RDF (View Source for full doctype...)>
- <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xml:base="http://www.mpii.de/yago/resource" xmlns:rdfs="http://www.w3.org/2000/01/rdf-
  schema#" xmlns:y="http://www.mpii.de/yago/resource/">
- <rdfs:Class rdf:about="http://www.mpii.de/yago/resource/wikicategory_Capitals_in_Asia">
  <rdfs:subClassOf rdf:ID="f600000009"
    rdf:resource="http://www.mpii.de/yago/resource/wordnet_capital_108518505" />
</rdfs:Class>
- <rdf:Description rdf:about="#f600000009">
  <y:confidence
    rdf:datatype="http://www.w3.org/2001/XMLSchema#double">0.9511911446218017</y:confidence>
</rdf:Description>
- <rdfs:Class rdf:about="http://www.mpii.de/yago/resource/wikicategory_Coastal_cities">
  <rdfs:subClassOf rdf:ID="f600000021"
    rdf:resource="http://www.mpii.de/yago/resource/wordnet_city_108524735" />
</rdfs:Class>
- <rdf:Description rdf:about="#f600000021">
  <y:confidence
    rdf:datatype="http://www.w3.org/2001/XMLSchema#double">0.9511911446218017</y:confidence>
</rdf:Description>
- <rdfs:Class
  rdf:about="http://www.mpii.de/yago/resource/wikicategory_Cities_towns_and_villages_in_the_United
  <rdfs:subClassOf rdf:ID="f600000033"
    rdf:resource="http://www.mpii.de/yago/resource/wordnet_town_108665504" />
</rdfs:Class>

```

Figure 8 YAGO: Snippet of subsumption hierarchy (taken from the YAGO project site)

Suchanek et al. implemented a query engine along the lines of [Kas08] on top of the database version of YAGO. The engine processes queries given in a SPARQL (<http://www.w3.org/TR/rdf-sparql-query/>) fashion, as shown in Table 10, by performing word resolution and then issuing a single SQL query that contains one `SELECT` argument for each variable to be bound and one join for each line of the query. (The details are described in [Suc08].) Suchanek et al. also provide a demonstrative search interface to query the YAGO knowledge base, which will be described in Section 4.3.

Table 10 YAGO: Sample queries (taken from [Suc07])

Query	Result
When was "Mostly Harmless" written? (<code>Mostly_Harmless, WRITTENINYEAR, \$y</code>)	<code>\$y=1992</code>
Which humanists were born in 1879? (<code>\$h, TYPE SUBCLASSOF*, humanist</code>) (<code>\$h, BORNINYEAR, 1879</code>)	<code>\$h=Albert_Einstein</code> and 2 more
Which locations in Texas and Illinois bear the same name? (<code>\$t, LOCATEDIN, Texas</code>) (<code>\$n, MEANS, \$t</code>) (<code>\$n, MEANS, \$k</code>) (<code>\$k, LOCATEDIN, Illinois</code>)	<code>\$n="Farmersville"</code> and 121 more

YAGO has been linked to other ontologies and datasets on the Web, including, e.g., DBpedia, SUMO (Suggested Upper Merged Ontology) (<http://www.ontologyportal.org>), and Freebase (<http://freebase.com>).

There are similarities and differences between the YAGO project and the PanAnthropon project, which can be discussed in terms of data models, data extraction approaches, and data storage schemes.

In contrast to the YAGO project, which is based on an extensive data model using the formalism of description logics, this research does not attempt at providing a strict model-theoretic framework for defining the semantics of the data elements and their relations. Nevertheless, the underlying data model in this research is quite similar overall to the one in YAGO, although there are certain differences.

In the data model of this research, things of all kinds can be considered as entities, but not necessarily so. That is, certain things, in particular, some of the literal values, both numeric and non-numeric, are simply left as literal values. Since such values are not considered as entities, there are no corresponding classes for such values. The decision on what is considered as an entity and what is not considered as such is not a rigidly fixed one, and is based on the consideration of the respective merit of deciding in one way or another for a given value type (within the given domain).

Such flexibility in the data model does not impede a consistent semantic representation of the data, since the data model of this research is based on the `<entity, attribute, value>` view of data, where `value` can be a class, entity, or literal. In a sense, the `<entity, attribute, value>` view places a restriction on the value type for the first argument, `entity`, which can only be a non-literal entity. This is in contrast to the `<entity, relation, entity>` view of YAGO, where there is no restriction on either of the two `entity` arguments (since everything, including a literal, is considered as an entity). However, the apparent restriction in the data model in this research is not a real restriction for this research, because the decision on which literal value type to consider as an entity or not is based precisely on the weighing of the relative utility of the given value type when it occupies the position of the (first) `entity` argument (i.e., the position of `subject` in the `<subject, predicate, object>` triple).

In the data model and the corresponding data storage scheme of this research, classes are considered and stored separately from other common entities, in such a way that the subsumption hierarchy of classes is stored in a separate database table (without explicitly using the attribute `subClassOf`) while facts concerning class membership of common entities (i.e., in the form of `<entity, type, class>` and `<entity, subtype, subclass>`) are stored in a different table containing all entity-centric facts. (Wikipedia categories are also considered as special types of entities, and are stored separately from other entities. The details of the data storage scheme and database structure will be discussed in Chapter 6.)

In contrast to the YAGO data model, the current data model in this research does not consider relations (i.e., attributes) as entities, and, accordingly, no properties pertaining to relations, or relations between relations, are explicitly extracted or recorded. However, this does not preclude implicitly exploiting some of the common-sense properties of relations. (For example, from a fact such as `<film directed_by director>` (where `film` and `director` are to be replaced by individual entity names), an inverse fact of `<director directed_film film>` can be derived, as will be described in Chapter 6.)

The data model of this research, like the YAGO data model, assigns a fact ID to each fact in the form of `<entity, attribute, value>`. As such, it is possible, in principle, to consider each such fact as an entity, and refer to it as such, by virtue of the unique ID assigned (which is called a reification process), as in YAGO. However, such has not been attempted, in part because the data model of this research has another mechanism to achieve the same purpose in the case of representing non-binary relations or of providing contextual information (namely, by adding the `note` field, as will be described in Chapter 5), and in part because this research either does not consider or differently processes the meta-relations considered in YAGO.

Even though both the YAGO project and the PanAnthropon project aim at extracting (and retrieving) (semi-)structured semantic/ontological information, a basic difference between the YAGO approach and that of this research consists in the fact that the former is concerned with general-domain information extraction/retrieval whereas the latter is geared toward domain-oriented information extraction/retrieval.

The stance of this research is that a domain-aware approach can still accommodate extensibility and inter-domain interoperability (if necessary), both on the conceptual/ontological level and on the practical level, while it can more effectively extract/retrieve/present information concerning a given domain, as will be discussed in Chapter 5 (and partly illustrated in Chapters 6 and 7).

Since it is geared toward open-domain information extraction, and since it aims at building a large-scale knowledge base containing a huge number of (heterogeneous) entities and facts, it is appropriate that YAGO used the dump of the entire English Wikipedia corpus. In contrast, this research used a relatively small, selected subset of Wikipedia (mainly 10,640 pages on films) as the major source of information.

On the surface, the number of entities (209,266) and the number of (entity-centric) facts (≈ 2.35 million) extracted/derived through this research pale in comparison to the figures given by YAGO, i.e., 1.7 million entities and 15 million facts (not to mention the current figures of 10 million entities and 80 million facts in YAGO2, which uses GeoNames (<http://www.geonames.org/>) in addition to Wikipedia and WordNet). However, such a face-value comparison does not put the matter into proper perspective. For, the English Wikipedia corpus used by Suchanek et al. in [Suc08] already contained 2 million article pages.

Such being the case, it turns out that, *on average*, the YAGO system extracted less than 1 entity per page and 10 facts per entity. (And this is the case, despite the fact that absolutely everything is considered as an entity according to the YAGO data model and that non-binary facts are represented with multiple triples.) In contrast, this research extracted/derived nearly 20 entities per page and more than 11 facts per entity.

From the proper comparison, therefore, it may be argued that the system of the PanAnthropon project was far more effective than the YAGO system in terms of extracting/deriving entities and facts, even by using a far smaller dataset. And this is precisely the intent of this research, namely, to demonstrate the ability to extract/derive a good (not necessarily huge) number of entities and facts of high quality that can be effectively used for entity retrieval concerning the chosen application domain.

The above comparison, of course, does not address the question of scalability or efficiency (in terms of the cost-benefit analysis). But investigating those issues is beyond the scope of this dissertation. Suffice it to say that the approach used in this research is intended to be extensible or adaptable to larger datasets and/or other domains and/or other data sources besides Wikipedia.

Apart from the difference in the size of the respective Wikipedia datasets used for information extraction, there is also a difference between the YAGO project and the PanAnthropon project in terms of which portions of Wikipedia pages were processed for the purpose. While this research also used infoboxes and categories as in the YAGO project, the former also processed, in addition, the abstract section and the cast member information section in the film pages, which do not have standardized structured formats and are therefore far more difficult to process in order to extract semantic information.

The methods used for information extraction/derivation and the types of information extracted/derived from infoboxes and categories are, however, similar between the YAGO project and this research, albeit not identical. Besides using infoboxes to extract `<attribute, value>` pairs, as in YAGO, this research also used heuristics similar to that used in the YAGO project to differentiate between different types of categories. The differences are that, in this research, some of the relational and thematic categories were converted to conceptual categories, that categories themselves are organized in a hierarchical taxonomy but are treated separately from classes and from ordinary entities, and that categories were used to derive mainly non-class entities (i.e., individuals) rather than classes, and, accordingly, non-taxonomic facts.

The `describes` relation in YAGO, which holds between a URL and the given individual, is not treated as a separate fact in this research, so that the URL is stored in a table that contains basic information on the entities. This research did not consider `foundIn` (between a fact and a URL) and `extractedBy` (between a fact and an extraction technique) relations in YAGO. As it did not extract hyperlinks as such, this research also did not extract facts involving an individual and the individuals connected via hyperlinks, which are represented by the `context` relation in YAGO.

Finally, unlike the YAGO project, which used plain text files as internal formats for storing the extracted information, this research used a MySQL database by using a combination of different data models and table schemas, as will be explained in Chapter 6.

DBpedia (<http://dbpedia.org>), much like YAGO, is the product of an effort to extract structured information from Wikipedia and make this information available on the Web, both for human- and machine-consumption, thereby making it possible to issue sophisticated queries against the extracted dataset as well as to link the latter with other datasets on the Web.

As shown in Figure 9, the DBpedia project consists of the following main components:

- Methods for extracting information from Wikipedia
- DBpedia dataset in RDF format
- Methods for accessing the DBpedia dataset
- Methods for interlinking DBpedia with other open datasets
- Interfaces for querying the DBpedia dataset

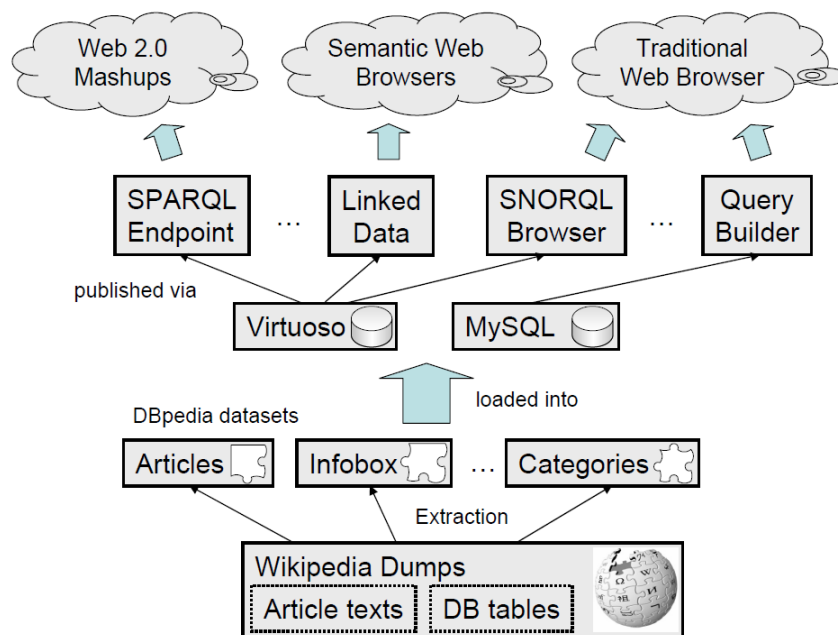


Figure 9 DBpedia: Overview of components (taken from [Auer07b])

The DBpedia components are described below in turn, with a focus on the information extraction methods and the resulting dataset, except the search interfaces, which will be described in Section 4.3.

In constructing the DBpedia dataset, Auer et al. used two different methods for extracting information: First, they used the SQL version of Wikipedia dumps (<http://dumps.wikimedia.org/>) in order to convert the semantic relationships already stored in relational database tables to triples

of the form `<subject, predicate, object>` (\approx `<entity, attribute, value>`) in the RDF (<http://www.w3.org/RDF/>) format. Second, they extracted additional semantic/ontological information from infobox templates and article texts, via pattern matching over the MediaWiki (<http://www.mediawiki.org>) markup.

The algorithm developed (and implemented in PHP) by Auer et al. to extract information from infobox templates, such as the one shown in Figure 10, proceeds in five stages:

- 1) Identification of Wikipedia pages containing templates
- 2) Extraction of significant templates
- 3) Parsing of each significant template
- 4) Post-processing of object values
- 5) Determination of the class membership for a given page

```
{{Infobox Town AT |
name = Innsbruck |
image_coa = InnsbruckWappen.png |
image_map = Karte-tirol-I.png |
state = [[Tyrol]] |
regbzk = [[Statutory city]] |
population = 117,342 |
population_as_of = 2006 |
pop_dens = 1,119 |
area = 104.91 |
elevation = 574 |
lat_deg = 47 |
lat_min = 16 |
lat_hem = N |
lon_deg = 11 |
lon_min = 23 |
lon_hem = E |
postal_code = 6010-6080 |
area_code = 0512 |
licence = I |
mayor = Hilde Zach |
website = [http://innsbruck.at] |
}}
```

Innsbruck	
	
Country	Austria
State	Tyrol
Administrative region	Statutory city
Population	117,342 (2006)
Area	104.91 km ²
Population density	1,119 /km ²
Elevation	574 m
Coordinates	47°16' N 11°23' E 
Postal code	6010-6080
Area code	0512
Licence plate code	I
Mayor	Hilde Zach
Website	www.innsbruck.at 

Figure 10 Wikipedia: MediaWiki markup of infobox template and its output (taken from [Auer07a])

Selecting all Wikipedia pages that contain templates: The algorithm retrieves Wikipedia pages with templates by using an SQL query searching for occurrences of the template delimiters “{{” and “}}” in the text table of the MediaWiki database layout.

Selecting and extracting significant templates: The algorithm extracts only those templates with a high probability of containing structured information, filtering out those that contain just one or two attributes and those whose usage count is below a set threshold.

Parsing each template and generating appropriate triples: The URL derived from the title of the Wikipedia page, in which the given template appears, is used as the subject of the triples. Each attribute is converted to the predicate, and the corresponding attribute value is converted to the object.

Post-processing object values to generate appropriate URI references or literal values: In case an attribute value is a linked object (e.g., “[[Troll]]” in line 4 of Figure 10), a suitable URI reference is generated, which refers to the linked Wikipedia page. For strings and numeric values, semantically-typed literals are generated by detecting and encoding common units as special data types, as shown in Table 11. In case an attribute value consists of a comma-separated list of multiple values, such a list may be converted to an RDF list or individual statements, depending on the configuration options.

Determining the class memberships for Wikipedia pages: Considering that the Wikipedia category structure does not constitute a strict subsumption hierarchy, as also recognized by Suchanek et al., Auer et al. mention that they are working to improve class membership detection (without explaining how).

Table 11 DBpedia: Examples of typed literals (taken from [Auer07a])

Attribute type	Example	Object data type	Object value
Integer	7,058	xsd:integer	7058
Decimals	13.3	xsd:decimal	13.3
Images	[[Image:Innsbruck.png 30px]]	Resource	c:Innsbruck.png
Links	[[Tyrol]]	Resource	w:Tyrol
Ranks	11th	u:rank	11
Dates	[[January 20]] [[2001]]	xsd:date	20010120
Money	\$30,579	u:Dollar	30579
Large numbers	1.13 [[million]]	xsd:Integer	1130000
Big money	\$1.13 [[million]]	u:Dollar	1130000
Percent values	1.8%	u:Percent	1.8
Units	73 g	u:Gramm	73

Even though Auer et al. provide a description of the method and process of extracting information from infoboxes, as summarized above, they do not provide a description concerning information extraction from article texts. As such, it is not clear what kind of method they used, if any, and what kind of information, if any, was extracted. It may be that Auer et al. simply extracted the abstract portions of the pages without further using them in order to extract semantic information.

According to the project Web site, the DBpedia knowledge base currently contains more than 3.5 million things (i.e., entities), out of which 1.67 million are classified in a consistent ontology, and over 672 million RDF triples (i.e., facts), including 286 million extracted from the English version of Wikipedia.

Figure 11 shows a snippet of the visual overview of the class subsumption hierarchy in the DBpedia ontology. (Note that, as in all OWL ontologies, `owl:Thing` is the root (or top-level) class.) Table 12 lists (in both columns) the names of only level-1 classes (directly subsumed by `owl:Thing`). (The complete list of classes, including all lower-level classes, can be found at <http://mappings.dbpedia.org/server/ontology/classes>.)

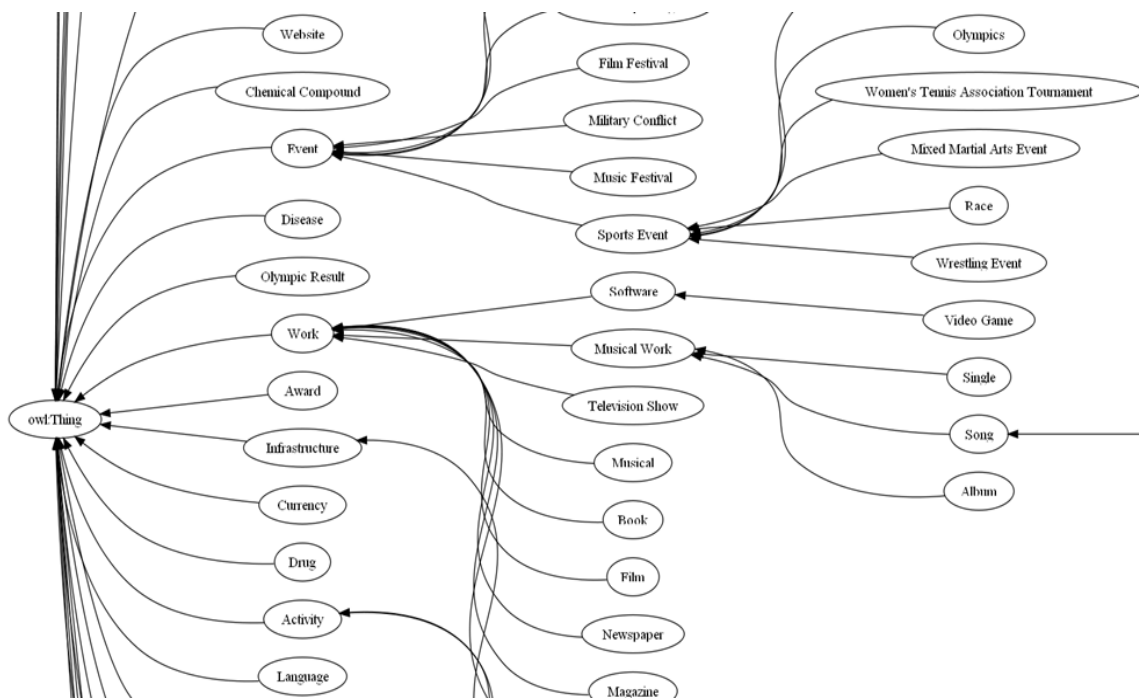


Figure 11 DBpedia: Snippet of ontology (taken from the DBpedia project site)

Table 12 DBpedia: Level-1 classes in ontology

Activity	LegalCase
AnatomicalStructure	MeanOfTransportation
Award	MusicGenre
Beverage	OlympicResult
ChemicalCompound	Organisation
Colour	Painting
Currency	Person
Device	PersonFunction
Disease	Place
Drug	Planet
EthnicGroup	Protein
Event	Sales
GovernmentType	Species
Infrastructure	Website
Language	Work

The DBpedia dataset consists of the component datasets shown in Table 13. Current statistics on the content of the DBpedia dataset, according to the statement on the DBpedia site, is shown in Table 14.

Table 13 DBpedia: Component datasets

Ontology Infoboxes	Article Categories	Redirects
Ontology Infobox Properties	External Links	Disambiguation Links
Ontology Infobox Properties (Specific)	Raw Infobox Properties	Categories (Labels)
Titles	Raw Infobox Property Definitions	Categories (Skos)
Short Abstracts	Homepages	Page IDs
Extended Abstracts	Geographic Coordinates	Revision IDs
Images	Pagelinks	PND
Links to Wikipedia Articles	Persondata	

Table 14 DBpedia: Current statistics on dataset content

persons	364,000
places	462,000
music albums	99,000
films	54,000
video games	17,000
organizations	148,000
species	169,000
diseases	5,200
links to images	1,850,000
external links to Web pages	5,900,000
external links into other RDF datasets	6,500,000
Wikipedia categories	633,000
YAGO categories	2,900,000

Auer et al. provide three mechanisms for accessing the DBpedia dataset: Linked Data [Ber06] [Biz07], SPARQL endpoint, and downloadable RDF dumps. For the details concerning these access mechanisms, the reader is directed to [Auer07b]. Here it may only be mentioned that DBpedia resource URIs (e.g., http://dbpedia.org/resource/The_Godfather) are set up to return a set of RDF descriptions, when accessed via Semantic Web agents, and an HTML view of the same information, when accessed via traditional Web browsers. The latter case is exemplified in the snippet shown in Figure 12. The DBpedia knowledge base is interlinked with various other datasets on the Web using RDF links, as illustrated in Figure 13, taken from the Linked Data (<http://linkeddata.org/>) project Web site.

dbpedia-owl:cinematography	■ dbpedia:Gordon_Willis
dbpedia-owl:director	■ dbpedia:Francis_Ford_Coppola
dbpedia-owl:distributor	■ dbpedia:Paramount_Pictures
dbpedia-owl:editing	■ dbpedia:Peter_Zinner ■ dbpedia:William_H._Reynolds
dbpedia-owl:gross	■ 1.33698921E8
dbpedia-owl:musicComposer	■ dbpedia:Carmine_Coppola ■ dbpedia:Nino_Rota
dbpedia-owl:producer	■ dbpedia:Albert_S._Ruddy
dbpedia-owl:starring	■ dbpedia:Talia_Shire ■ dbpedia:Richard_Conte ■ dbpedia:Marlon_Brando ■ dbpedia:Diane_Keaton ■ dbpedia:John_Cazale ■ dbpedia:Sterling_Hayden ■ dbpedia:Richard_S._Castellano ■ dbpedia:John_Marley ■ dbpedia:Al_Pacino ■ dbpedia:Abe_Vigoda ■ dbpedia:Robert_Duvall ■ dbpedia:Al_Lettieri ■ dbpedia:Gianni_Russo ■ dbpedia:James_Caan_(actor)
dbpedia-owl:subsequentWork	■ dbpedia:The_Godfather_Part_II

Figure 12 DBpedia: Sample resource accessed via regular Web browser

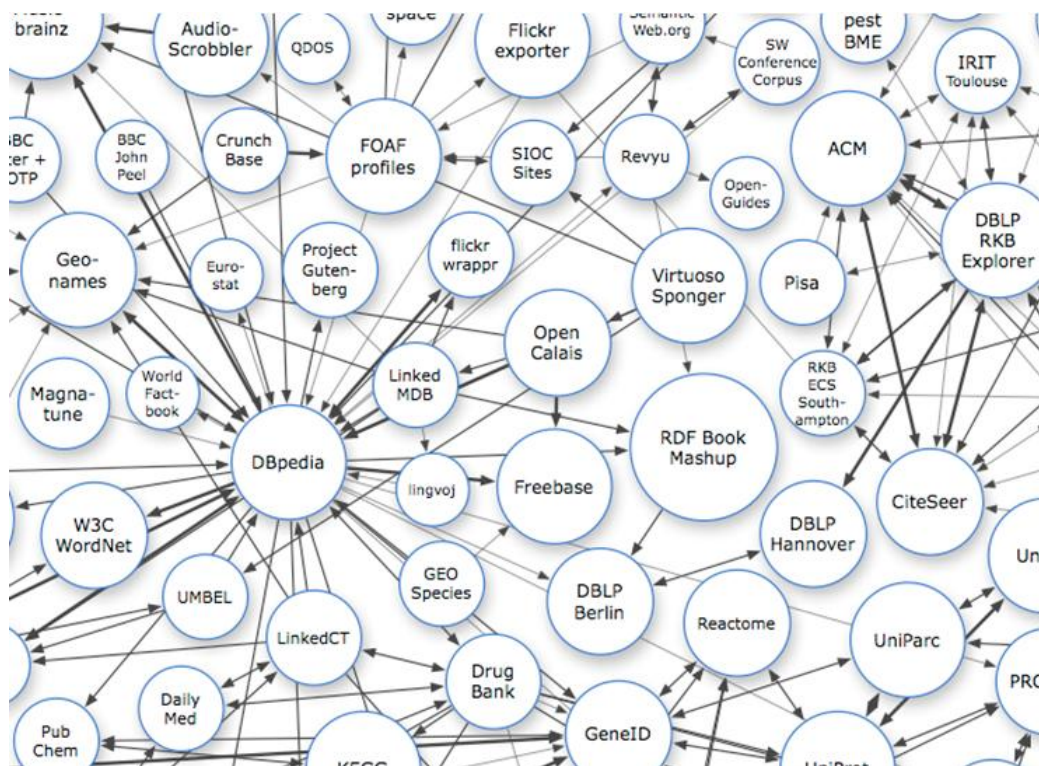


Figure 13 DBpedia: Interlinked data resources (taken from the Linked Data site)

The DBpedia project is quite similar to the YAGO project in intent, purpose, and methodology. As such, in general, pretty much similar things can be said about the DBpedia project, as about the YAGO project, in contrast to the PanAnthropon project. Namely, such as the fact that the DBpedia project is concerned with general- or multi-domain information extraction, that it used a much larger Wikipedia corpus as the source dataset, that it extracted information using Wikipedia pages rendered in MediaWiki markup (in contrast to this research that parsed Wikipedia pages in HTML format), that it mainly used only infoboxes and categories to extract semantic information, that it uses a knowledge representation formalism for the Semantic Web (i.e., RDF) to represent/record the extracted data, that, as such, it is accessible by Semantic Web agents and is interlinked with many other Semantic Web resources, etc.

The last two points mentioned above, about data representation/storage and about data linking, have bearings upon this research in terms of possible directions of future work.

First, as to data representation/storage, although the entities and entity facts extracted through this research are currently represented/stored as tuples in (relational) tables in a MySQL database, the data can be converted to other representation formalisms or formats, including RDF and XML, although such conversion is outside the scope of this dissertation. Since the data are represented using a semantic data model, which is only slightly different from the `<subject, predicate, object>` model of RDF, it is expected that such conversion will not pose a big problem.

Second, as to data linking, once the dataset of this research is converted to RDF, for example, it will be possible to link the dataset to other semantic data resources, including, foremost, DBpedia and YAGO, which are already interlinked with each other. Such data linking will require resolving the URIs in order to discover and link the same entities in different datasets by using

the `owl:sameAs` property. In this regard, it is also possible to link the dataset of this research to the DBpedia/YAGO datasets indirectly, by linking the former to a dataset that is already linked to DBpedia/YAGO, such as LinkedMDB.

Although it is not concerned with extracting information (directly) from Wikipedia, the Linked Movie Database (LinkedMDB) (<http://linkedmdb.org>) project [Has09] is related to the PanAnthropon project in terms of the domain of application, as it is related (and linked) to the DBpedia and YAGO projects in terms of the goal of creating semantically-interlinked datasets.

The goal of the LinkedMDB project is to create the first open linked dataset connecting movie-related resources on the Web, as illustrated in Figure 14. As the main data source, the project used Freebase (<http://www.freebase.com>), a collaboratively-built database which contains information on various topics (including 38,000+ films), much of the information coming from Wikipedia.



Figure 14 LinkedMDB: Sample entities (taken from [Has09])

LinkedMDB provides links to datasets, including DBpedia and YAGO, within the Linking Open Data (LOD) (<http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData>) cloud. It also provides links to external Web pages of other movie-related Web sites such as IMDb (The Internet Movie Database) (<http://www.imdb.com/>), omdb (Open Media Database) (<http://www.omdb.org/movie>), and Rotten Tomatoes (<http://www.rottentomatoes.com/>). Table 15 shows the overall statistics on the LinkedMDB dataset. Table 16 shows statistics on sample entity types in the dataset. Table 17 shows interlinking statistics.

Table 15 LinkedMDB: Overall statistics on dataset content (taken from [Has09])

Total number of triples	3,579,616
Number of interlinks to LOD cloud	162,199
Number of links to movie websites	271,671
Number of entities in LinkedMDB ¹	233,103

Table 16 LinkedMDB: Statistics on sample entity types (taken from [Has09])

Entity	Count
Film	38,064
Actor	29,361
Director	8,367
Writer	12,990
Producer	9,637
Music Contributor	3,995
Cinematographer	2,169
Interlink	162,199

Table 17 LinkedMDB: Statistics on interlinked resources (taken from [Has09])

Target	Type	Count
DBpedia	<code>owl:sameAs</code>	30,354
YAGO	<code>owl:sameAs</code>	30,354
flickr wrappr	<code>DBpedia:hasPhotoCollection</code>	30,354
RDF Book Mashup (Books)	<code>movie:relatedBook</code>	700
RDF Book Mashup (Authors)	<code>rdfs:seeAlso</code>	12,990
MusicBrainz	<code>owl:sameAs</code>	2,207
GeoNames	<code>foaf:based_near</code>	27,272
GeoNames	<code>owl:sameAs</code>	272
lingvoj	<code>movie:language</code>	28,253
IMDb, RottenTomatoes Freebase.com	<code>foaf:page</code>	271,671

Similarly as in YAGO and DBpedia, LinkedMDB is accessible via traditional Web browsers, Semantic Web browsers, and SPARQL clients. Figure 15 shows a snippet of information on the film *Patton* (<http://data.linkedmdb.org/resource/film/291>), accessed via a regular Web browser, which is quite similar to the snippet shown in Figure 12.

Property	Value
movie:actor	< http://data.linkedmdb.org/resource/actor/30085 >
movie:actor	< http://data.linkedmdb.org/resource/actor/31792 >
movie:actor	< http://data.linkedmdb.org/resource/actor/37477 >
movie:actor	< http://data.linkedmdb.org/resource/actor/38034 >
movie:actor	< http://data.linkedmdb.org/resource/actor/40885 >
movie:actor	< http://data.linkedmdb.org/resource/actor/40908 >
movie:actor	< http://data.linkedmdb.org/resource/actor/432 >
movie:actor	< http://data.linkedmdb.org/resource/actor/51744 >
movie:actor	< http://data.linkedmdb.org/resource/actor/52343 >
movie:actor	< http://data.linkedmdb.org/resource/actor/52945 >
movie:actor	< http://data.linkedmdb.org/resource/actor/53284 >
movie:actor	< http://data.linkedmdb.org/resource/actor/54345 >
movie:actor	< http://data.linkedmdb.org/resource/actor/65609 >
movie:actor	< http://data.linkedmdb.org/resource/actor/65627 >
movie:actor	< http://data.linkedmdb.org/resource/actor/9676 >
movie:cinematographer	< http://data.linkedmdb.org/resource/cinematographer/51 >
dc:date	1970-02-04
movie:director	< http://data.linkedmdb.org/resource/director/8412 >
movie:editor	< http://data.linkedmdb.org/resource/editor/3084 >
movie:film_cut	< http://data.linkedmdb.org/resource/film_cut/4812 >
movie:film_format	< http://data.linkedmdb.org/resource/film_format/61 >
is movie:film_of_distributor of	< http://data.linkedmdb.org/resource/film_film_distributor_relationship/252 >
movie:film_story_contributor	< http://data.linkedmdb.org/resource/film_story_contributor/342 >

Figure 15 LinkedMDB: Sample resource accessed via regular Web browser

From looking at the number of entities (233,103) in Table 15 and the number of films (38,064) in Table 16, a similar remark can be made about the relative effectiveness of entity extraction, when comparing the LinkedMDB project and the PanAnthropon project, as when comparing the latter with the YAGO project. For, it shows that the LinkedMDB project extracted 6 entities per film, compared to 18~20 entities (depending on whether or not to count the films that do not have Wikipedia pages) in this research.

Moreover, it is quite interesting to compare the number of entities in the LinkedMDB dataset vs. the PanAnthropon FilmWorld dataset, as shown in Table 18.

Table 18 LinkedMDB vs. PanAnthropon: Comparison of number of entities

Entity (Sub)Type	LinkedMDB	PanAnthropon
film	38,064	11,355
actor	29,361	45,615
director	8,367	4,184
writer	12,990	10,209
producer	9,637	7,460
musician (music contributor)	3,995	3,321
cinematographer	2,169	2,344

As shown, this research extracted only slightly smaller numbers of writers, producers, and musicians, compared to those in LinkedMDB, despite the fact that the size of the source dataset (i.e., the number of films) used in this research is less than a third of that in LinkedMDB. The number of directors in this research database is about a half the number in LinkedMDB, which is fully expected. In contrast, the number of cinematographers is slightly larger in the dataset of this research. Furthermore, the number of actors extracted through this research is 1.5 times larger than that in LinkedMDB. In particular, the figures indicate that the LinkedMDB project extracted less than 1 (≈ 0.77) actor per film, on average, whereas this research extracted at least 4 (≈ 4.02) actors per film. (It may be noted that there are partial overlaps between the numbers of entities for different person entity subtypes shown for this research, because a person may assume multiple roles. However, there should be similar overlaps in the LinkedMDB dataset.)

Given that both projects are concerned with the film domain and that LinkedMDB is already linked with various resources, it would be appropriate to consider the possibility of linking the PanAnthropon FilmWorld dataset with the LinkedMDB dataset. However, this issue is beyond the scope of this dissertation and is left as a topic for possible future work.

4.3 Information Retrieval on Wikipedia Data

Despite its prominence as a foremost information source on the Web, Wikipedia provides a basic search interface that is limited to entering keywords and retrieving a list of articles whose titles and/or content (partially) match the keywords entered by the user (unless an article whose title exactly matches the query string exists). Correspondingly, various systems have been developed for the sake of providing alternative methods for searching the content of Wikipedia [Cat08], none of which, however, is geared toward the task of finding entities that directly match a query.

Suppose we want to find films directed by Werner Herzog, which belong to the genre of epic film, which were released in the 1970s or 1980s. How could we find such films in Wikipedia? We may enter a keyword query as follows: `film Werner Herzog epic 1970s 1980s`. Figure 16 shows the top 5 out of 13 article pages returned. (It is peculiar that only 13 pages have been returned. The titles of the remaining 8 article pages are: “Film”, “Ambient music”, “Akira Kurosawa”, “Silent film”, “Cinema of Iran”, “Wilhelm von Homburg”, “Isabella Adjani”, and “Vampire”.) As shown in Figure 16 and as suggested by their titles, the pages returned are at best somehow related to Werner Herzog and at worst purely coincidental. In all cases, the matching of a page with the query is based on checking the occurrences of individual keywords. None of the 13 pages has directly to do with the films we are looking for. One may consider that such poor search results may have resulted from the fact that too little context was provided in the query. Accordingly, we may try modifying the query as follows: `epic films directed by Werner Herzog released in 1970s 1980s`. Or, we may even try: `films directed by Werner Herzog, which belong to the genre of epic film, and which were released in the 1970s or 1980s`. The result, however, is that we get shorter lists of pages (9 in the first case, 3 in the second case), which nevertheless consist of the subsets of the same 13 pages.

Search results

From Wikipedia, the free encyclopedia

[Content pages](#) [Multimedia](#) [Help and Project pages](#) [Everything](#) [Advanced](#)

[Art film](#) (section [1970s](#))

An **art film** (also known as "art movie", "art house **film**", or in the collective ... cinema s, or in the US "arthouse cinemas") and **film** festivals

48 KB (7,094 words) - 16:20, 6 June 2010

[History of film](#) (section [1980s: sequels, blockbusters and videotape](#))

The history of **film** is the historical development of the medium known variously as ... early **1970s**, English- ... cinema, with **Werner Herzog** , Rainer ...

160 KB (24,743 words) - 18:41, 9 June 2010

[Cinema of Germany](#) (redirect from [German film](#))

Schlöndorff , **Werner Herzog** , Jean-Marie ... existing German **film** industry and ... prominence in the **1980s**.The ... Programmkinos) from the **1970s** onwards ...

42 KB (6,187 words) - 22:32, 17 May 2010

[List of films shot in Thailand](#) (section [1980s](#))

Dozens of foreign **films** have been shot in Thailand, with the kingdom either playing ... The availability of elephant s, exotic jungle and beach ...

19 KB (2,659 words) - 01:49, 9 May 2010

[1970s in film](#) (section [S](#))

Figure 16 Wikipedia: Sample search result

Alternatively, we may try finding the films we are looking for from the Wikipedia article page on Werner Herzog, which contains the filmography section as shown in Figure 17. From the list of feature films directed by Werner Herzog, we can identify those that were released in the 1970s or 1980s. However, we cannot identify which films belong to the genre of epic film unless we check Wikipedia article pages corresponding to each film. Still another way in which we may try finding the films we are looking for is to take advantage of the Wikipedia category page on “Films directed by Werner Herzog”, as shown in Figure 18. In this case, however, it will be even harder to find the films that match the specified conditions, since we will have to check both the film release year and the film genre for each film on the list. Trying to initiate the search from the “1970s in film” or “1980s in film” pages will, of course, further broaden the scope of the search and make it even harder to find the films sought.

Filmography

All films were directed and written (or co-written) by Werner Herzog:

Features

- *Signs of Life* (1968)
- *Even Dwarfs Started Small* (1970)
- *Fata Morgana* (1971)
- *Aguirre, the Wrath of God* (1972)
- *The Enigma of Kaspar Hauser* (1974)
- *Heart of Glass* (1976)
- *Stroszek* (1977)
- *Nosferatu the Vampyre* (1979)
- *Woyzeck* (1979)
- *Fitzcarraldo* (1982)
- *Where the Green Ants Dream* (1984)
- *Cobra Verde* (1987)
- *Scream of Stone* (1991)
- *Invincible* (2001)
- *The Wild Blue Yonder* (2005)
- *Rescue Dawn* (2007)
- *Bad Lieutenant: Port of Call New Orleans* (2009)
- *My Son, My Son, What Have Ye Done* (2009)
- *The Piano Tuner* (201?)

Figure 17 Wikipedia: Filmography section of page on Werner Herzog

Category:Films directed by Werner Herzog

From Wikipedia, the free encyclopedia

For more information, see [Werner Herzog](#).

Pages in category "Films directed by Werner Herzog"

The following 53 pages are in this category, out of 53 total. This list may not reflect recent changes ([learn more](#)).

A	H	P cont.
• Aguirre, the Wrath of God	• Handicapped Future	• Precautions Against Fanatics
B	• Heart of Glass (film)	R
• Bad Lieutenant: Port of Call New Orleans	• Herakles (film)	• Rescue Dawn
• Ballad of the Little Soldier	• Werner Herzog	S
• Bells from the Deep	• How Much Wood Would a Woodchuck Chuck (film)	• Scream of Stone
C	• Huie's Sermon	• Signs of Life (1968 film)
• Cobra Verde	I	• La Soufrière (film)
D	• Invincible (2001 film)	• Stroszek
• The Dark Glow of the Mountains	J	T
• Death for Five Voices	• Jag Mandir (film)	• Ten Minutes Older
E	L	• Ten Thousand Years Older
• Echoes from a Somber Empire	• Land of Silence and Darkness	• The Transformation of the World into Music
• Fata Morgana	• Last Words (film)	U
• Fitzcarraldo	• Last Words (film)	
• Fata Morgana		
• Where the Green Ants Dream		
• Cobra Verde		
• Scream of Stone		
• Invincible		
• The Wild Blue Yonder		
• Rescue Dawn		
• Bad Lieutenant: Port of Call New Orleans		
• My Son, My Son, What Have Ye Done		
• The Piano Tuner		

Figure 18 Wikipedia: Category page on films directed by Werner Herzog

Wikiwix (<http://www.wikiwix.com/>) provides a single interface for simultaneously searching content across English Wikipedia, Wiktionary, Wikisource, etc., as shown in Figure 19. The list of Wikipedia pages returned for the sample query *for a sample run* includes 29 pages titled “Werner Herzog”, “1970s in film”, “Aguirre, the Wrath of God”, “Epic film”, “Vampire films”, etc. The list includes a film that matches the specified conditions — *Aguirre, the Wrath of God* — among the top 3 results, as shown in Figure 20, even though the only indication that the page may be potentially relevant to the query is given by the presence of the word “Herzog” displayed in bold. Interestingly, when the exactly same query was entered again, the result was different, which suggests that the above result was obtained only by chance.



Figure 19 Wikiwix: Search interface

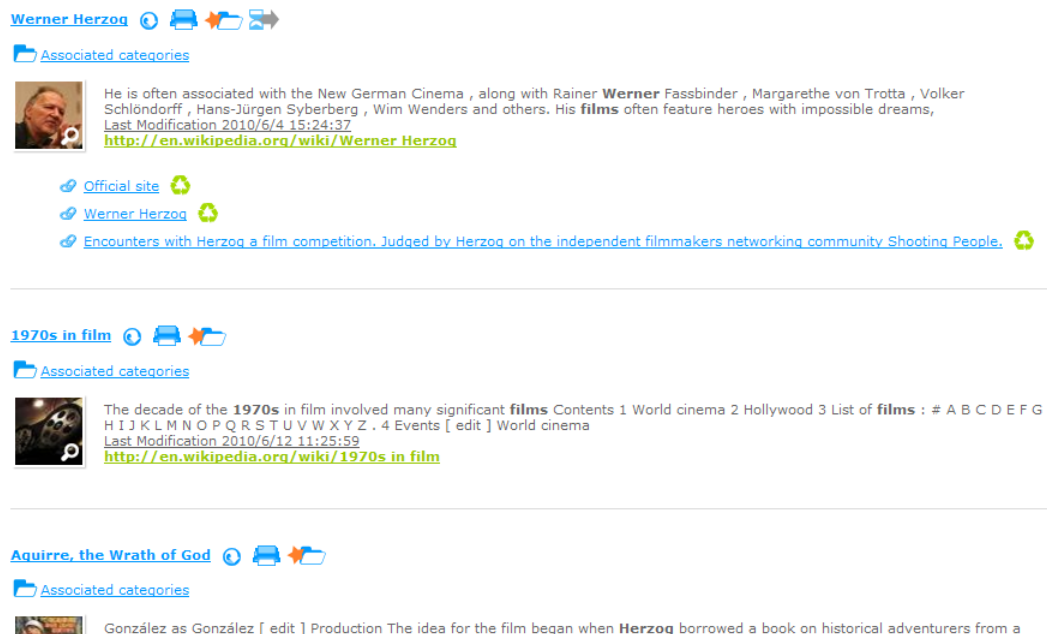


Figure 20 Wikiwix: Sample search result

Figure 21 shows the top 5 Wikipedia pages returned by Powerset (<http://www.powerset.com/>) upon entering the same query. Similar to the results returned by Wikiwix, the list includes the film page “Aguirre, the Wrath of God” among the top 5 results. In the case of Powerset search results, however, we get slightly better indications as to the relevance of that particular page to the query entered, given the phrases “Films directed by Werner Herzog”, and “Epic films” (both corresponding to Wikipedia categories associated with the page) being highlighted. Nevertheless, as in Wikipedia and Wikiwix, the query–result matching is based on keyword matching. Consequently, pages that have only general relevance to the query, such as “1970s in film” and “Epic film”, are ranked similarly to those that have more specific relevance, such as “Aguirre, the Wrath of God” and “Werner Herzog”. This is of course due to the fact that keyword matching does not take into account semantic relevance but only textual relevance.

The screenshot shows the Powerset search interface. At the top, the search bar contains the query "Epic films directed by Werner Herzog released in 1970s 1980s" and a "search" button. Below the search bar, there are options for "Wikipedia Articles", "hide highlighting", "advanced", and a help icon. The search results are displayed as a list of four items, each with a dropdown arrow on the left:

- Werner Herzog** | | Film chronology · German Empire 1895–1918 · Weimar Germany 1919–1933 · Nazi Germany 1933–1945 · East Germany (1945–1990) · (West) Germany 1945–present · 1945–1959 · 1960s · **1970s** · **1980s** · 1990s · 2000s · 2010s Actors · Directors · **Films A–Z** · Cinematographers · Festivals · Producers · Composers · Screenwriters | | ... **Films directed by Werner Herzog**
- Epic film** **Werner Herzog's** 'Aguirre: The Wrath of God' didn't cost as much as the catering in 'Pearl Harbor,' but it is an **epic**, and 'Pearl Harbor' is not." ... **Epic films**
- 1970s in film** These included **directors** like Wim Wenders, Hans-Jürgen Syberberg and **Werner Herzog**. ... The end of the decade saw two **epic** Vietnam War **films**, from directors Michael Cimino (The Deer Hunter) and Coppola (Apocalypse Now).
- Aguirre, the Wrath of God** **Films directed by Werner Herzog** ... **Epic films**
- My Best Fiend** | | Film chronology · German Empire 1895–1918 · Weimar Germany 1919–1933 · Nazi Germany 1933–1945 · East Germany (1945–1990) · (West) Germany 1945–present · 1945–1959 · 1960s · **1970s** · **1980s** · 1990s · 2000s · 2010s Actors · Directors · **Films A–Z** · Cinematographers · Festivals · Producers · Composers · Screenwriters | | ... **Films directed by Werner Herzog**

Figure 21 Powerset: Sample search result

The other Wikipedia search systems on Catone's list [Cat08] are farther from the task of finding entities and are intended for various other purposes. Similpedia (<http://www.similpedia.org/>), for example, takes a URL or a paragraph of text as input and returns a list of English Wikipedia article pages that potentially have similar content to the input text/page. Figure 22 shows the search interface. Figure 23 shows sample search results given the "Aguirre, the Wrath of God" film page URL. Note that the input page itself is returned at the top of the list of similar pages.

Figure 22 Similpedia: Search interface

Similpedia Results - Similar Content in Wikipedia

Aguirre, the Wrath of God

Aguirre, the Wrath of God Aguirre, the Wrath of God (German : Aguirre, der Zorn Gottes) is an independent 1972 German film written and directed by Werner Herzog . Klaus Kinski stars in the title role. The soundtrack was composed and performed by German progressive/ Krautrock band Popol Vuh. Aguirre was given an extensive arthouse theatrical release in the United States in 1977, and remains one of the director's most famous films. The story follows the travels of Spanish soldier Lope de Aguirre, who leads a g...

http://en.wikipedia.org/wiki/Aguirre,_the_Wrath_of_God

Portrait Werner Herzog

Portrait Werner Herzog Portrait Werner Herzog () is an autobiographical short film by Werner Herzog made in 1986. Herzog tells stories about his life and career. The film contains excerpts and commentary on several Herzog films, including Signs of Life , Heart of Glass, Fata Morgana , Aguirre, the Wrath of God, The Great Ecstasy of Woodcarver Steiner , Fitzcarraldo, and the Les Blank documentary Burden of Dreams. Notable is footage of a conversation between Herzog and his mentor Lotte Eisner , and a discuss....

http://en.wikipedia.org/wiki/Portrait_Werner_Herzog

My Best Fiend

My Best Fiend My Best Fiend (German: Mein liebster Feind - Klaus Kinski, literally My Dearest Enemy - Klaus Kinski) is a 1999 documentary film by Werner Herzog about his tumultuous yet productive relationship with German actor Klaus Kinski. It was released on DVD in 2000 by Anchor Bay. Summary The film opens with shots from one of Klaus Kinski 's Jesus tours, in which he performed – after his own interpretation – the role of Jesus. Kinski harangues the audience for not paying attention to him, curses wi...

http://en.wikipedia.org/wiki/My_Best_Fiend

Aguirre (disambiguation)

Aguirre (disambiguation) Aguirre is a Basque (Spanish) surname. Aguirre may also refer to * Aguirre, the Wrath of God , 1972 German film about Lope de Aguirre, written and directed by Werner Herzog ** Aguirre (soundtrack) , the soundtrack to Herzog's film, performed by Popol Vuh * Aguirre Department , Santiago del Estero Province, Argentina * BAP Aguirre, the name of several Peruvian Navy ships commissioned between 1951 and 2005 {{disambig}}....

[http://en.wikipedia.org/wiki/Aguirre_\(disambiguation\)](http://en.wikipedia.org/wiki/Aguirre_(disambiguation))

Werner Herzog

Figure 23 Similpedia: Sample search result

WikiWax (<http://www.wikiwax.com/>) provides a dynamic list of search suggestions while the user is typing keywords to find a Wikipedia article. As shown in Figure 24 and Figure 25 below, the mechanism is only geared toward finding potential Wikipedia article titles, not toward finding entities or even finding information on pages. Upon user selection, the system simply directs the user to the selected article page.

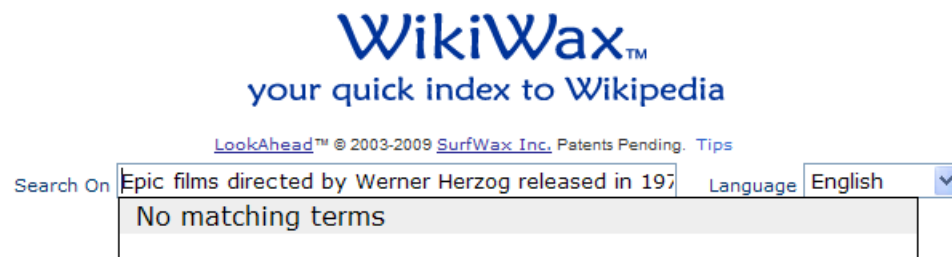


Figure 24 WikiWax: Search suggestion — no matching titles found



Figure 25 WikiWax: Search suggestion — numerous matching titles found

WikiMindMap (<http://www.wikimindmap.org/>) is an information visualization tool that presents a Wikipedia article, whose title exactly matches the query string, in the interactive visual form of a mind map consisting of branches representing the section titles and leaves representing the hyperlinks that appear in a given section, as shown in Figure 26. Even though the tool can help one get a quick (and nice) overview of a given page and navigate within the page or to the related pages connected via hyperlinks, the system does not have search functionalities beyond these.

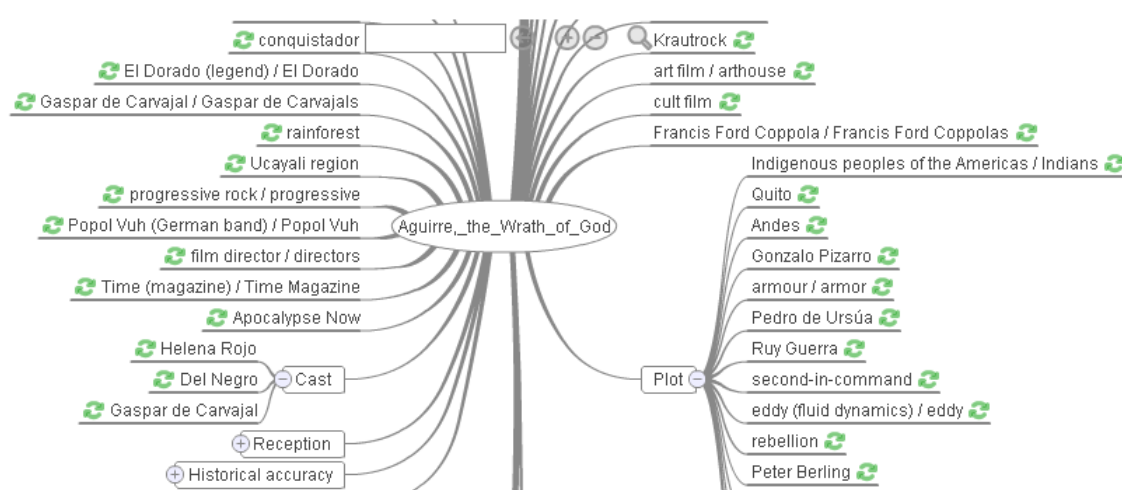


Figure 26 WikiMindMap: Sample search result

Among researchers who worked on extracting lexical or semantic information from Wikipedia, Milne et al. [Mil07b], Suchanek et al. [Suc07] [Suc08], and Auer et al. [Auer07a] [Auer07b], for example, presented search systems/interfaces based on the data extracted from Wikipedia.

Based on a thesaurus they constructed using Wikipedia, Milne et al. [Mil07b] built a search engine, called Koru, that assists the users in the process of finding Wikipedia articles, by providing relevant query topic candidates that match the keywords entered by the users and showing the corresponding query results, thereby helping the users modify their queries so as to find the articles that better match their intentions.

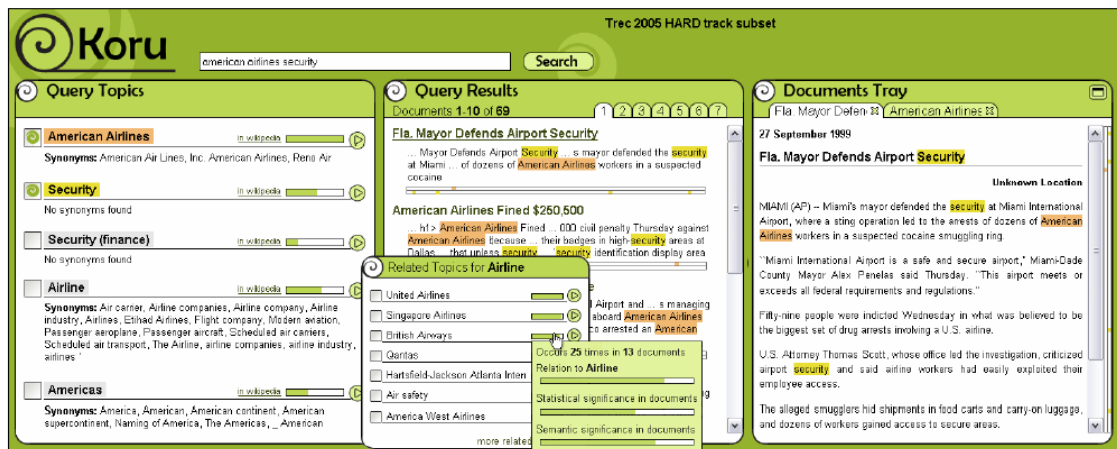


Figure 27 Koru: Search interface with sample search result (taken from [Mil07b])

As shown in Figure 27, the interface consists of three panels, which are selectively presented to the user as the search process progresses. The first stage of the search process is concerned with building a query, which involves adding/removing phrases until the query and corresponding search results satisfy the user's information need. At this stage, the leftmost two panels, query topics and query results, are visible to the user. The second stage is concerned with browsing the list of documents returned so as to determine the most relevant one. At this stage, the rightmost two panels, query results and document tray, are visible. The final stage is concerned with reading the selected document, with only the document tray panel being visible. While the query topic suggestion functionality can be helpful for finding topically relevant Wikipedia article pages, the system's search functionality does not go beyond keyword-based document retrieval.

YAGO of Suchanek et al. [Suc07] [Suc08] and DBpedia of Auer et al. [Auer07a] [Auer07b] provide interfaces for querying the semantic information extracted from Wikipedia by using SPARQL query patterns, consisting of a set of conditions, each in the form of <subject, predicate, object>. In contrast to all the other systems reviewed in this section, the YAGO and DBpedia interfaces are geared toward finding specific entities (or facts concerning entities).

Figure 28 shows the query pattern examples provided on the YAGO query demo page (<http://www.mpi-inf.mpg.de/yago-naga/yago/demo.html>). As shown, the user can input the query by specifying up to four conditions to be satisfied by the entities (or facts) sought. The fields corresponding to `subject` and `object` can each be filled with an entity individual name (with initial capitalization), a class (in lowercase), or a variable (prefixed by `?`). The field corresponding to `predicate` can be filled with a specific predicate or an open predicate (marked by `?`).

Querying

For demonstration purposes, we provide an interface for querying YAGO in a SPARQL-like fashion below.

- Give me people called "Fabian" (Try your own name!)

Fabian	givenNameOf	?whom	try
--------	-------------	-------	---------------------

- Tell me everything you know about Angela Merkel (relation is a questionmark)

Angela Merkel	?	?what	try
---------------	---	-------	---------------------

- Give me scientists who born in Ulm (two conditions)

?x	isA	scientist	try
?x	bornIn	Ulm	

- Which Nobel prize winners were born after Albert Einstein? (many conditions)

Albert Einstein	bornOnDate	?EinsteinDate	try
?x	bornOnDate	?otherDate	
?otherDate	after	?EinsteinDate	
?x	hasWonPrize	Nobel Prize	

Figure 28 YAGO: Query examples (taken from the YAGO project site)

As shown in Figure 29, the actual query form provides a dropdown menu containing all available predicates to choose from. Since YAGO is a general-domain knowledge base, and since the query form does not impose any restrictions as to what types of entities can occupy the `subject` field, the dropdown menu contains all predicates, regardless of whether or not a given predicate may be relevant to the entity in the `subject` field. Figure 30 shows top 2 results returned for the query looking for films directed by Werner Herzog and actors who acted in those films.

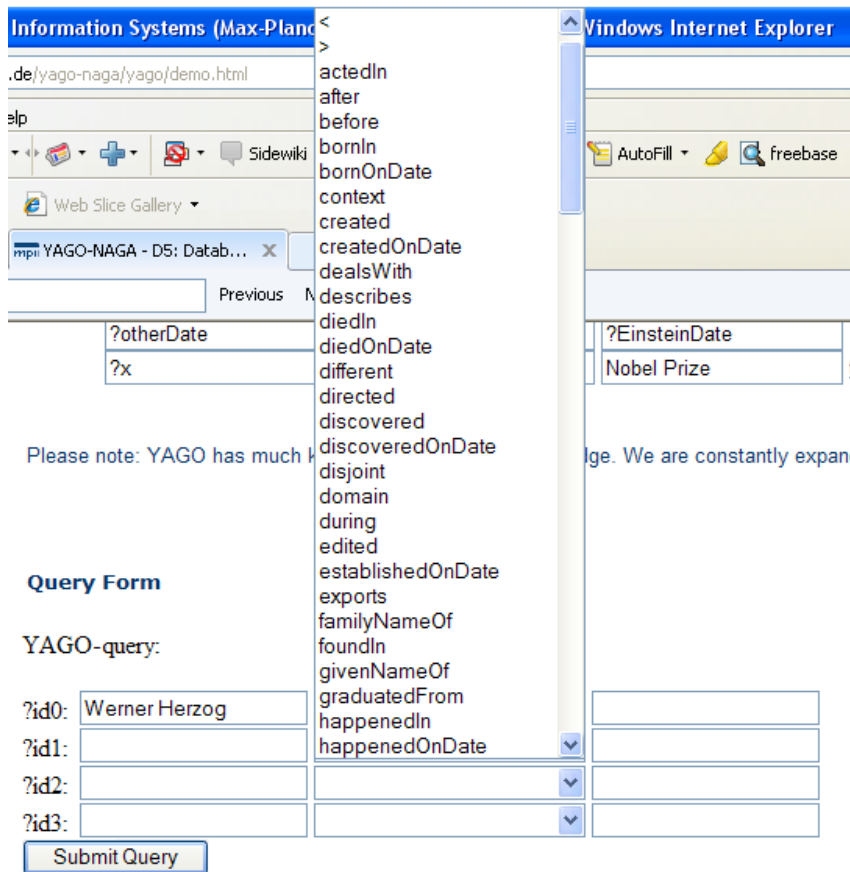


Figure 29 YAGO: Dropdown menu containing predicates

Query Form

YAGO-query:

?id0:	Werner Herzog	directed	?x
?id1:	?y	actedIn	?x
?id2:			
?id3:			

Submit Query

?Werner Herzog = [Werner Herzog](#)
 ?x = [Lessons of Darkness](#)
 ?y = [Werner Herzog](#)

?Werner Herzog = [Werner Herzog](#)
 ?x = [The Wild Blue Yonder](#)
 ?y = [Brad Dourif](#)

Figure 30 YAGO: Sample query result

Figure 31 shows the DBpedia query builder interface, as presented in [Auer07a]. (At the time of this writing, the query builder on the DBpedia Web site (<http://querybuilder.dbpedia.org/>) was inaccessible due to the redesign process in progress.) Except the fact that the predicate field here provides suggestions for predicates using the look-ahead technology, instead of providing a dropdown menu containing available predicates, and except the fact that query results are presented in a table format rather than in a list format, the query form and query input format are quite similar to those of the YAGO interface.

Wikipedia Query Builder

Query: Please provide triple patterns, the results should match to. Prefix variables with "?", use ">", "<", "=", "~" (Regex) for comparisons. Alternatives can be given by using "|".

Subject	Predicate	Object
<input type="text" value="?city"/>	<input type="text" value="leader_name"/>	<input type="text" value="?leader"/>
<input type="text" value="?city"/>	<input type="text" value="subdivision_name"/>	<input type="text" value="United_States"/>
<input type="text" value="?city"/>	<input type="text" value="e"/>	<input ">1000"="" type="text" value=""/>

Results

Click on a column header to sort this page: 10

?city	?leader	>1000
Cadillac, Michigan	Ronald Blanchard	1328
Denver, Colorado	John Hickenlooper (D)	1609
Roswell, New Mexico	Sam LaGrone	1089
Santa Fe, New Mexico	David Coss	2231
Las Cruces, New Mexico	William Michael Mattiace	4000 ft - 1219
Artesia, New Mexico	Manuel Madrid	1030
Carlsbad, New Mexico	Robert Forrest	1004

Download

NTriples dump of all extracted RDF statements from Wikipedia: [wikipedia.nt.bz2](http://wikipedia.3ba.se/film/wikipedia.nt.bz2)

Sourcecode is available from: <http://powl.sf.net>

Save current query

Label Save

Previously saved queries

- [Soccer player with tricoot number 11 from club with stadium with >40000 seats born in a country with more than 10M inhabitants](#)
- [Films with music from John Williams](#)
- [Films with Quentin Tarantino as actor, producer, or director](#)

Figure 31 DBpedia: Query Builder interface (taken from [Auer07a])

The search interface created from this research is similar to those of YAGO and DBpedia, both in terms of purpose and in terms of appearance, insofar as it is geared toward finding entities/facts rather than pages and insofar as it also employs a form-based query input method for specifying semantic conditions. There are, however, distinct differences, as will be described in Chapter 7.

CHAPTER 5: CONCEPTUALIZATION

5.1 Conceptual Stance

In this research, “entities” are conceived of as things of all kinds, both concrete and abstract, that have certain “attributes” (or properties), either independent or relational, either inherent or assigned. The `type` (attribute) of an entity refers to — more precisely, relates the entity to — a generic “class” into which the given entity is classified, e.g., `person`, `work`, etc. In general, the `type` of an entity — the class that corresponds to the value of the attribute `type` for the given entity — is fixed and exclusive in the sense that an entity that belongs to one class does not or cannot belong to other classes. The `subtype` of an entity refers to a subclass into which the entity can be classified, under a given class. The `subtype` of an entity can be fluid and non-exclusive in the sense that an entity may belong to more than one subclass, under a given class. This is especially so in the case of `person-type` entities, and thus a `subtype` may better be understood as a *role* in this case. In general, there are multiple subclasses under a given class, and the former can be further classified into still more specific subclasses. A “fact” refers to a tuple in the form of `<entity, attribute, value, note>`, which adds the `note` argument to the `<entity, attribute, value>` triple model. The value of an attribute can be (an instance of) a literal, an entity, a class, or a (Wikipedia) category. In case an attribute mediates two entities (i.e., in case another entity corresponds to the value for the attribute), such an attribute is considered as representing a “relation” between two entities. The class membership relation is a special kind of relation that holds between an entity and a class/subclass, mediated by the attribute `type/subtype`. The kinds of classes/subclasses and attributes/relations that are relevant, except certain basic kinds, depend on the domain at issue (i.e., the universe of discourse). An entity may belong to (or may be related to) multiple domains, but not every class, attribute, or relation involving the entity is relevant or equally important in one domain as in another domain.

5.2 Ontological Scheme

The purpose of an ontological scheme is to define, coherently and consistently, the kinds of classes into which entities are to be classified. The implication of the conceptual stance laid out in Section 5.1 is that the task of constructing such a scheme can better be approached from the viewpoint of a given domain of application which effectively constitutes the ontological space in which the entities are located. Such domain-oriented ontological schemes need not be mutually exclusive or incompatible. They can accommodate extensibility or inter-domain interoperability.

Based on the view as stated above, this research takes a domain-oriented approach to ontology construction as well as to information extraction/retrieval. In this research, the process of constructing an ontology and the process of extracting/deriving information mutually informed each other, so that new classes were added, when applicable, as the latter process progressed.

Table 19 shows the film-domain-oriented ontology constructed from the FilmWorld version of the PanAnthropon project. (The ontology is presented in a space-economical tabular format in order to provide a better overview of the structure of the entire ontology while keeping the class names legible.) Each column of the table corresponds to a different level in the subsumption hierarchy in a descending order from left to right, starting from the top level to level 5. (For example, the following subsumption relations hold: `director` is a subclass of `film artist`, which is a subclass of `artist`, which is a subclass of `person`, which is a subclass of `thing`.) To clarify the meaning of some of the class names in the ontology: The class `place` refers to geographical unit. Similarly, the class `time` refers to temporal unit or interval. It is noted that the line between `cultural convention` and `cultural artifact` is not necessarily clear-cut. Nevertheless, the two were distinguished in order to indicate the distinction of event-related vs. static nature.

Table 19 PanAnthropon: Film-domain-oriented ontology

Top	Level 1	Level 2	Level 3	Level 4	Level 5
thing	person	artist	film artist	director	
				actor	
				cinematographer	
				animator	
				editor	
			narrator		
		musician			
		writer			
		entrepreneur	film entrepreneur	producer	
				distributor	
	work	work of art	film		
	organization	company	film company		
			music company		
			other company		
		government agency			
	other organization				
	place (geo unit)	continent			
		country			
		constituent country			
		state/province/region			
		city			
	time (tempo unit)	century			
		decade			
		year			
		year-month			
		month			
		month-day			
	date				
	event	award event	film award event		
	cultural convention	award	film award		
	cultural artifact	language			
	concept	artwork-related concept	film-related concept	film (acting) role	
				film genre	
				film theme/subgenre	
				film subject	generic subject
					specific subject
				film setting	temporal setting
					geographical setting
				film (plot) source	generic source
				specific source	
		film type			
	film style				
person-related concept	person-name-related concept	given name			
		family name			
technology	artwork-related technology	film (making) technology			

The ontology shown in Table 19 was intended to be lightweight and extensible. Lightweight, in the sense that only those classes that are necessary or useful for the purpose of retrieving information on the film domain in this research are included in the ontology. Extensible, in the sense that more classes (whether those concerning the film domain or some other domain, or

those that are domain-independent) can be incorporated without (or at least without much) restructuring of the ontology.

When an ontology, such as the one shown in Table 19, is presented, the usual questions raised are about the validity and/or completeness of the given ontology, as if there were definitive methods to judge each and every ontology, beyond any reasonable doubt, as either valid or invalid, either complete or incomplete. Indeed, the very fact that such questions are raised indicates that no such methods are known to the questioners, for, otherwise, they would have already known the answers, without the need of raising the questions. Suppose the questions are not about binary decisions but are instead about the matters of degree. Still, how could one measure/show how (much) valid/invalid or how (much) complete/incomplete a given ontology is? Usually, however, the questions about the validity and/or completeness of a given ontology are raised in the manner of demanding the source or basis of the ontology, i.e., where it comes from, what (other ontology) it is based on. Even if any such one (or two or many) source(s) could be provided as the basis of the given ontology, how would one then know whether or not *that* source ontology itself is valid and/or complete? If, now, the question about the source/basis should be raised about that source ontology itself, such a questioning would lead (at least theoretically) to quasi-infinite regression, unless and until an “unmovable mover” kind of source ontology were to be found. Since we can safely assume, in this particular kind of world of all possible worlds, that there would be only a finite number of ontologies in the series of ontologies basing themselves on other ontologies, we can also safely expect to arrive at the final element of the series at some point in time. Suppose we did. How could we know whether or not *this* ontology is valid and/or complete? We could not. We could either accept it as is, in which case the ontology we started with would be automatically justified, *or* reject it as baseless, in which case no ontology in the series would be deemed valid/complete.

Philosophical reflections aside, let us here consider the DBpedia ontology for reference purposes. Table 20 (continued on pp.77–78) shows level-1 and level-2 classes in the DBpedia ontology. (The entire subsumption hierarchy (<http://mappings.dbpedia.org/server/ontology/classes>) also consists of 5 levels, as in the PanAnthropon FilmWorld ontology. But it is not necessary for our present purposes to look at all the lower-level classes in the ontology.)

Table 20 DBpedia: Upper-level classes in ontology

Top	Level 1	Level 2
Thing	Activity	Game
		Sport
	AnatomicalStructure	Artery
		Bone
		Brain
		Embryology
		Lymph
		Muscle
		Nerve
		Vein
	Award	
	Beverage	
	ChemicalCompound	
	Colour	
	Currency	
	Device	AutomobileEngine
		Weapon
	Disease	
	Drug	
	EthnicGroup	
	Event	Convention
		FilmFestival
		MilitaryConflict
		MusicFestival
		SpaceMission
		SportsEvent
	YearInSpaceflight	
	GovernmentType	
	Infrastructure	Road
	Language	
	LegalCase	SupremeCourtOfTheUnitedStatesCase

Table 20 (continued)

MeanOfTransportation	Aircraft
	Automobile
	Rocket
	Ship
	SpaceShuttle
	SpaceStation
	Spacecraft
MusicGenre	
OlympicResult	
Organisation	Band
	Broadcast
	Company
	EducationalInstitution
	GeopoliticalOrganisation
	Legislature
	MilitaryUnit
	Non-ProfitOrganisation
	PoliticalParty
	RadioStation
	SportsLeague
	SportsTeam
	TradeUnion
Painting	
Person	Ambassador
	Architect
	Artist
	Astronaut
	Athlete
	BritishRoyalty
	Celebrity
	Cleric
	CollegeCoach
	Criminal
	FictionalCharacter
	Journalist
	Judge
	MilitaryPerson
	Model
	Monarch
	OfficeHolder
	Philosopher
	PlayboyPlaymate
	PokerPlayer
	Politician
	Scientist
	SoccerManager

Table 20 (continued)

	PersonFunction	
	Place	BodyOfWater
		Building
		Cave
		HistoricPlace
		LunarCalendar
		Monument
		Mountain
		MountainPass
		MountainRange
		Park
		PopulatedPlace
		ProtectedArea
		SiteOfSpecialScientificInterest
		SkiArea
		Valley
		WineRegion
	WorldHeritageSite	
	Planet	
	Protein	
	Sales	
	Species	Archaea
		Bacteria
		Eukaryote
	Website	
	Work	Book
		Film
		Magazine
		Musical
		MusicalWork
		Newspaper
		Software
		TelevisionEpisode
		TelevisionShow

Let it be noted that *this* ontology (along with the YAGO ontology) is one of *the* most authoritative ontologies within the Semantic Web community. Let it also be reminded that *this* is the “consistent” ontology according to which at least 1.67 million entities are classified in the DBpedia dataset. Furthermore, the entities in numerous semantic data resources linked to/with DBpedia are linked to/with (some of) those DBpedia entities classified under *this* ontology.

Now, first, let us look at level-1 classes. We notice that those classes not only include generic classes, e.g., `Activity`, `Event`, `Organisation`, `Person`, `Place`, `Work`, etc., as expected, given that level-1 classes are those immediately subsumed by the top class `Thing`, but also rather specific classes (classes that refer to more specific kinds of entities), which would have been more appropriate to be placed at lower levels in the class hierarchy, e.g., `Award`, `Beverage`, `MusicGenre`, `OlympicResult`, `Painting`, `Protein`, `Sales`, `Website`, etc. Given that the first level of class hierarchy effectively stipulates the types of entities to be placed at all lower levels, such inconsistency within level-1 classes has adverse ramifications for the entire ontology.

If we now also look at level-2 classes, we notice that inconsistency exists, not only within a given level of the class hierarchy, but also between levels, i.e., between a class and its subclasses. For example, the direct subclasses of the class `Person` include `BritishRoyalty`, `CollegeCoach`, `FictionalCharacter`, `PlayboyPlaymate`, `SoccerManager`, etc., alongside (more generic kinds such as) `Artist`, `Politician`, `Scientist`, etc. One should wonder what kind of rationale lies behind classifying, e.g., `PlayboyPlaymate`, as an immediate subclass of `Person`. One would also wonder if `FictionalCharacter` is a kind of person at all. Such inconsistency is not limited to the `Person` class and not limited to only level-2 classes. For example, although the DBpedia ontology includes `Film` (level 2) as a subclass of `Work` (level 1), and `Actor` (level 3) as a subclass of `Artist` (level 2) (which is a subclass of `Person`), it does not have a class corresponding to `Director` anywhere within the ontology.

Given the observations above, one could hardly argue that the DBpedia ontology is valid and complete, in general, and that it is more valid and more complete than the ontology in this research, in particular. In any case, it is not the intention of this research to argue for the validity/completeness of its ontology that is only created/used as a tool to serve its purpose.

5.3 Semantic Entity Typing

The entities extracted/derived through the information extraction process described in Chapter 6 are semantically typed (i.e., classified) according to the ontology shown in Table 19. In particular, the `type` of an entity refers to a level-1 class, whereas the `subtype` of an entity refers to a leaf class subsumed by the former, as shown in Table 21 (continued on p.81). While this is the internal structure used to classify entities, a simplified scheme of entity type/subtype presentation, shown in Table 22, is used for the menu options in the search interface described in Chapter 7.

Table 21 PanAnthropon: Entity types/subtypes

Entity Type	Entity Subtype
person	director
	actor
	cinematographer
	animator
	editor
	narrator
	musician
	writer
	producer
	distributor
work	film
organization	film company
	music company
	other company
	government agency
	other organization
place	continent
	country
	constituent country
	state/province/region
	city
time	century
	decade
	year
	year-month
	month
	month-day
	date
event	film award event
cultural convention	film award
cultural artifact	language

Table 21 (continued)

concept	film (acting) role
	film genre
	film subgenre/theme
	film subject — generic
	film subject — specific
	film setting — temporal
	film setting — geographical
	film (plot) source — generic
	film (plot) source — specific
	film type
	film style
	given name
	family name
technology	film (making) technology

Table 22 PanAnthropon: Simplified entity types/subtypes

Entity Type	Entity Subtype
person	(any)
work	film
organization	(any)
place (geographical unit)	continent
	country
	constituent country
	state/province/region
	city
time (temporal unit)	century
	decade
	year
	year-month
	month
	month-day
event	date
cultural convention	film award event
cultural artifact	film award
cultural artifact	language
concept	film (acting) role
	film genre
	film theme/subgenre
	film subject — generic
	film subject — specific
	film setting — temporal
	film setting — geographical
	film (plot) source — generic
	film (plot) source — specific
	film type
	film style
	given name
	family name
technology	film (making) technology

CHAPTER 6: IMPLEMENTATION: INFORMATION EXTRACTION

The system was implemented by using Java servlets, Tomcat server, and MySQL database.

6.1 Information Extraction Process

6.1.1 Information Source

The first task for information extraction was to decide on the subset of English Wikipedia pages on films to be used as the main source of information. For this purpose, I first used the Wikipedia category page on “Years in film” (http://en.wikipedia.org/wiki/Category:Years_in_film) in order to extract the titles/URLs of 120 pages corresponding to each year in film history (e.g., “1950 in film”) between years 1890 and 2009, inclusive. I subsequently downloaded the 120 Wikipedia pages (in HTML).

From each page in the 120-page set, I then extracted the titles/URLs of films released in each year, by using the “Films released in xxxx” section as shown in Figure 32. (The actual section titles include variations such as: “Films released in xxxx”, “Notable Films released in xxxx”, “Notable films released in xxxx”, “Films”, “Wide releases”, “Wide-release films”, “Wide-release movies”, “Other films released”, etc.) (The film titles shown in red in Figure 32 represent the films for which no Wikipedia pages exist yet.)

A total of 11,355 film titles (and URLs) were extracted from the film release sections of the 120 pages. (The number 11,355 represents the number of films in the final dataset. A small number of films were filtered out at this stage or later for various reasons.) Each film in the 11,355-film set was considered as an entity, and was entered into the MySQL database, with information on the title, URL, entity type, entity subtype, and release year (as well as alternative page title and URL

in the case of a redirect page). (Information on the director and starring actors, when available for a no-page film, was also saved.) Wikipedia pages for the 10,640 films that have corresponding Wikipedia pages were subsequently downloaded and served as the main source of information.

Films released in 1918

- [Alraune](#)
- [Alraune, die Henkerstochter, genannt die rote Hanne](#)
- [Amarilly of Clothes-Line Alley](#), directed by Marshall Neilan, starring Mary Pickford
- [L'ame du bronze](#)
- [America's Answer](#)
- [Among the Cannibal Isles of the South Pacific](#)
- [The Bell Boy](#), a 'Fatty' Arbuckle / Buster Keaton short
- [Bound In Morocco](#), starring Douglas Fairbanks
- [The City of Dim Faces](#), starring Sessue Hayakawa and Marin Sais
- [The Cook](#), a 'Fatty' Arbuckle/Buster Keaton short
- [A Dog's Life](#), a Charlie Chaplin short
- [The Embarrassment of Riches](#)
- [Father Sergius](#), directed by Yakov Protazanov, starring Ivan Mozzukhin
- [The Goddess of Lost Lake](#), starring Louise Glaum
- [Headin' South](#), directed by Allan Dwan, starring Douglas Fairbanks
- [The Heart of Humanity](#) (starring Erich von Stroheim)
- [Hearts of the World](#), directed by D.W. Griffith, starring Lillian Gish and Dorothy Gish
- [The House of Mirth](#), directed by Albert Capellani, starring Katherine Corri Harris
- [Masks and Faces](#)
- [Mickey](#), directed by F. Richard Jones, starring Mabel Normand
- [Moonshine](#), a 'Fatty' Arbuckle/Buster Keaton short
- [Nocturnal Tunes](#)
- [Out of the Inkwell](#), animated film, directed by Max Fleischer
- [Out West](#), a 'Fatty' Arbuckle/Buster Keaton short

Figure 32 Wikipedia: Sample film release section

In addition to the film information pages, Wikipedia pages concerning two well-known film awards, i.e., Academy Awards and Golden Globe Awards, were also downloaded so that information on the award winners and nominees for each year (up to year 2010) of the award ceremonies could be extracted. (At the current stage of this research, information on only selected

award categories — Best Picture, Best Director, Best Actor, Best Actress, Best Supporting Actor, Best Supporting Actress, and Best Foreign Film — has been considered, since they represent the film award categories that are of most interest to general users. Information on the remaining award categories could of course be extracted and saved later on. Information on other prestigious film awards, e.g., Cannes Film Festival awards, etc. could also be added to the dataset.)

6.1.2 Direct Extraction of Information

For efficient processing, relevant sections of the downloaded film pages were separately stored for each film in the database, and information extraction was done by retrieving and processing those sections separately. In this way, processing of a given section was done for all films at once before processing of another section was started. The page sections used for information extraction include (the first paragraph of) the article abstract, the infobox, the categories, and the film cast information section.

The article abstract section was used to extract information on the films and to provide brief introductory excerpt for each film via the “Slide Show” function of the search interface described in Chapter 7. (The abstract was saved both in HTML format and in plain text format so that the former could be used for information extraction while the latter could be used for information presentation.) The abstract section was used to extract information on alternative titles of a film (i.e., `also_known_as` information). It was also used (after processing the infobox section and the film cast information section) to extract information on the film director(s), producer(s), writer(s), starring actors, and cast (and roles) in case a given film page does not contain an infobox, in case the infobox for a given film does not contain information fields for director, producer, writer, or starring actors, or in case a film page does not contain a film cast section.

Figure 33 shows a sample abstract taken from a film page. As shown, the abstract contains information on the alternate film titles as well as on the director, writers, and starring actors.

The Devil and Daniel Webster (film)

From Wikipedia, the free encyclopedia

The Devil and Daniel Webster is a 1941 fantasy film, adapted by Stephen Vincent Benét and Dan Tothoroh from Benét's short story, "The Devil and Daniel Webster". The film's title was changed to *All That Money Can Buy* to avoid confusion with another film released by RKO that year, *The Devil and Miss Jones*, and later had the title restored on some prints. It has also been released under the titles *Mr. Scratch*, *Daniel and the Devil* and *Here Is a Man*. The film stars Edward Arnold, Walter Huston, and James Craig. It was directed by William Dieterle.

Figure 33 Wikipedia: Sample film page abstract section

The abstract shown in Figure 33 is an example of relatively clearly-written, one-paragraph abstract. The length, content, and writing style of the abstract section vary widely among the film pages. Extracting information from the natural-language text in any case is quite challenging due to numerous variations for expressing the same thing. For example, the `also_known_as` facts were extracted by taking into account the cue words/phrases shown in Table 23, which can appear with or without parentheses/commas and in various combinations. Since some of the alternate film titles represent translations or transliterations of the original titles, or the titles by which a film is known in certain countries, an additional set of cue words/phrases was used to cover such cases. In addition, a set of stop words/phrases was used to prevent extraction of false positives (e.g., a film made by a director “known for” some other films).

Table 23 PanAnthropon: Cue words/phrases for extraction of `also_known_as` facts

aka	common	known	released	styled	translated	full title
a.k.a.	commonly	titled	re-released	stylized	tr.	onscreen title
also	frequently	re-titled	re-issued	spelled	transliterated	working title
or	often	retitled	reissued	written	translit.	under the title
alternate	sometimes		renamed	pronounced	literally	which means
alternately	original		premiered	abbreviated	lit.	
alternative	originally		marketed			
alternatively	officially					

The infobox section of a Wikipedia page contains multiple information fields consisting of `<attribute, value>` pairs. Not all Wikipedia pages have infoboxes. The attributes that appear in the infobox naturally differ among different domains and different types of entities, since the relevance of a given attribute depends on the domain and entity type. Even in the case of the same entity type within the same domain, a given page may contain more or less information fields than another one (due to the lack of information or due to the different stages of article development). In the case of film pages, the infoboxes can also differ according to the film genre and series. For example, some animated film page infoboxes contain attributes specific to the genre, such as “Animation by”, “Voices by”, “Layouts by”, “Backgrounds by”, etc.; James Bond series film page infoboxes contain different/additional information fields compared to other films.

Figure 34 shows a sample infobox taken from a film page. Even though infoboxes have semi-standardized formats, it is not in the least easy or straightforward to extract information from them. This is so, because, appearances to the contrary, there are numerous variations in which people enter the same information. These variations exist, not only among different pages, but also on the same page, within the same section, and even within the same information field in an infobox. In the case of infoboxes, for example, there are variations of font styles used to represent information field (attribute) names, subfield names, and value names, respectively, variations of ways in which multiple values for the same information field are delimited, variations of ways in which the same value is represented, etc. It has also been observed that some people enter “See below”, “See the article”, “See the xxx section”, etc. as attribute values.

Another difficult problem for information extraction, which is common to all Wikipedia page sections, not just the infobox section, is concerned with disambiguating entities in case an incorrect link is provided or in case no link is provided. There are a few causes of this problem.

Directed by	Alex Proyas
Produced by	Alex Proyas Andrew Mason
Written by	Screenplay: Alex Proyas David S. Goyer Lem Dobbs Story: Alex Proyas
Starring	Rufus Sewell Kiefer Sutherland Jennifer Connelly William Hurt
Music by	Trevor Jones
Cinematography	Dariusz Wolski
Editing by	Dov Hoenig
Distributed by	New Line Cinema
Release date(s)	February 27, 1993
Running time	100 minutes (theatrical cut) 111 minutes (director's cut)
Country	United States
Language	English
Budget	\$27 million
Gross revenue	\$27,200,316

Figure 34 Wikipedia: Sample film infobox section

One reason is simply that some Wikipedia contributors just do not seem to pay much attention to ensuring the quality of their contributions, even to the basic things such as making sure that correct links are provided as intended. (Interestingly, a certain demographic group of users has been (indirectly) observed to be the principal source of most formatting variations and lousy editorial styles, given that film pages pertaining to a film genre related to the ethnic group exhibit such patterns, almost without an exception.)

Another reason is that Wikipedia's editorial guidelines only require that at least the first occurrence of the mention of another Wikipedia article page need be hyperlinked, when applicable, not that all occurrences should be marked with hyperlinks. (The example of this practice is shown in the infobox in Figure 34. The name "Alex Proyas" is hyperlinked in the first, "Directed by" information field, but not in the subsequent fields.) This can be problematic, given the fact that the title of a hyperlinked page can provide disambiguating information that is absent in the plain text, e.g., John Smith (actor), John Smith (actor born in 1930), John Smith (American director), John Smith (British director), and John Smith (British director and writer).

Still another reason is that Wikipedia pages tend to be frequently removed and re-titled, instead of being redirected, thereby creating numerous, unintended bad links. For example, a film page originally titled "Amazing Adventures of John Smith" may be re-titled "Amazing Adventures of John Smith (film)" and another page concerning the source book upon which the film is based may be created with the original page title and URL. Later on, the "Amazing Adventures of John Smith (film)" page may become a disambiguation page which contains links to "Amazing Adventures of John Smith (1975 film)" and "Amazing Adventures of John Smith (1990 film)". In this commonly-observed scenario, links pointing to the first two pages can both become wrong links, after the fact.

Although the disambiguation problem is an important problem, it is out of the scope of this research to focus on devising an algorithm to tackle this issue, which, it is suspected, can have only limited effectiveness in solving the problem due to the different causes behind the problem. Within this research, the problem of duplication or confounding of entities due to the absence of links or the presence of bad links has been dealt with on a case-by-case basis, when such cases were brought to attention.

Information extraction from the infobox section was done for each information field at a time for all films. Table 24 shows the attributes that have been used in this research to extract information on the films. Not all attributes found in film page infoboxes have been used. Attributes such as “Budget”, “Gross revenue”, “Preceded by”, “Followed by”, and some genre-specific attributes have been intentionally excluded.

As shown in Table 24, the attributes in the infobox section relate a given film entity to other entities. For example, in the case of the attribute “Produced by” (`produced_by`), the corresponding value can consist either of an entity of type `person` (subtype: `producer`) or of an entity of type `organization` (subtype: `film company`). (Some values of inappropriate entity kinds for the given attribute (e.g., `animal` for the attribute “Starring” (`starring`), `person` for the attribute “Studio” (`from_studio`), etc.) were excluded.)

Not all infobox attributes have entities as their values. (Or, not all attribute values are considered as entities.) In the case of the attribute “Running time” (`has_running_time`), the corresponding value (e.g., “155 min.”, converted to “02:35:00” to enable comparison with other values) consists of (and is considered as) a literal. While it is possible to consider a value such as “02:35:00” as an entity instance of type `time duration` (subtype: `hr-min-sec`), it was decided not to do so, given its limited utility.


Besides the value for a given attribute, some context information (e.g., title of the original source book given along with the name of a writer for the attribute “Written by” (`written_by`)) was also extracted.

Table 24 PanAnthropon: Film attributes extracted from infobox

Original Attribute	Attribute	Value Type	Value Subtype
Directed by	directed_by	person	director
Produced by	produced_by	person	producer
		organization	film company
Written by Story by Novel/story by Screenplay by	written_by	person	writer
Narrated by	narrated_by	person	narrator
Starring James Bond Also Starring Voices by	starring	person	actor
Music by Composer Performer	music_by	person	musician
Cinematography	cinematography_by	person	cinematographer
Animation by	animation_by	person	animator
Editing by	editing_by	person	editor
		organization	film company
Studio	from_studio	organization	film company
Distributed by	distributed_by	organization	film company
		person	distributor
Release date(s)	released_in_year	time	year
	released_in_year-month		year-month
	released_in_month		month
	released_on_month-day		month-day
	released_on_date		date
Running time	has_running_time_of	[literal]	
Country	from_(or_co-produced_in)_country	place	country
Language	in_language	cultural artifact	language

Information concerning each film attribute was processed as follows: In case an entity serves as the value for an attribute, it was first checked whether or not the database already contained the entity. If not, information concerning the entity (i.e., entity name, entity page URL, entity type, and entity subtype) was first entered into the database. Information concerning the relationship between the film entity and the value entity at issue, mediated by the given attribute, was then entered into the database (along with additional context information). In case an attribute has a literal as its value, only the second type of information was entered into the database.

The categories section of a Wikipedia page contains special links that associate the given page with one or more Wikipedia categories, as shown in Figure 35.



Categories: [1990s action films](#) | [Action thriller films](#) | [1995 films](#) | [Films set in Pittsburgh, Pennsylvania](#) | [Sports in Pittsburgh, Pennsylvania](#) | [Ice Penguins](#)

Figure 35 Wikipedia: Sample film categories section

The Wikipedia category structure is not a tree structure, wherein a subcategory is subsumed by a single super-category, but rather a graph structure, wherein a subcategory can have multiple super-categories. (For example, “Action thriller films” is a subcategory of both “Action films” and “Thriller films”.) Moreover, there is no restriction on the level of categories to be associated with a given page, i.e., a page may be associated with both a super-category and its subcategories. (For example, a page may include category links to “Action films” and to “Action thriller films”.)

As in YAGO, administrative categories were excluded in this research. Unlike in YAGO, however, some relational categories and thematic categories were converted to conceptual categories (e.g., “Plays by Bertold Brecht” → Films based on plays by Bertold Brecht; “Macbeth on screen” → Films adapted from Macbeth; “Science fantasy” → Science fantasy films) while retaining the information on the original category names.

Categories were processed as follows: First, each unique category extracted was stored in the database (separately from entities) with a unique ID. Next, those categories were classified using “super-categories” (described below). Then information on the entity–category association was entered into the database for each film by using the attribute `associated_with_category`. Later on, the categories stored in the database were used also for indirect information derivation, based on their classification, as will be explained in Subsection 6.1.3.

A total of 5,229 unique categories were extracted from 10,640 film pages. Based on the consideration of the content of these categories, a hierarchical taxonomy consisting of a total of 215 super-categories was constructed, as shown in Table 25 (continued on pp.93–96). The taxonomy was stored in a separate table, apart from the table containing regular categories.

(Due to space limitation, the super-categories are shown only up level 3. The Film categories by year category has level-4 subcategories, e.g., 1890s film categories. The Film categories by director category also has subcategories: Film categories by director role/status/ethnicity and Film categories by director name.)

Unlike in Wikipedia, only one leaf super-category was assigned to a given regular category. (For example, Action thriller films was classified under the super-category Action film categories under Film categories by genre under Film categories.)

The purpose of such classification is twofold: First, the taxonomy itself is used to facilitate browsing the films by categories (via the “Category-Based Entity Browsing” function of the interface). Second, it was also used to derive additional classes, entities, and attributes to be used for the main entity search task.

Table 25 PanAnthropon: Taxonomy of super-categories

Level 1	Level 2	Level 3
Film categories	Film categories by timeline	Film categories by century
		Film categories by decade
		Film categories by year
	Film categories by country	Afghan film categories
		Algerian film categories
		American film categories
		Argentine film categories
		Australian film categories
		Austrian film categories
		Belgian film categories

Table 25 (continued)

	Bhutanese film categories
	Bosnia and Herzegovina film categories
	Botswana film categories
	Brazilian film categories
	British film categories
	Bulgarian film categories
	Burkinabé film categories
	Cameroonian film categories
	Canadian film categories
	Chilean film categories
	Chinese film categories
	Colombian film categories
	Croatian film categories
	Cuban film categories
	Cypriot film categories
	Czech film categories
	Czechoslovak film categories
	Danish film categories
	Democratic Republic of the Congo film categories
	Dutch film categories
	Egyptian film categories
	Fiji film categories
	Finnish film categories
	French film categories
	Georgian film categories
	German film categories
	Greek film categories
	Guinean film categories
	Hong Kong film categories
	Hungarian film categories
	Icelandic film categories
	Indian film categories
	Iranian film categories
	Irish film categories
	Israeli film categories
	Italian film categories
	Ivorian film categories
	Japanese film categories
	Kazakhstani film categories
	Lebanese film categories
	Libyan film categories
	Luxembourgian film categories
	Macedonian film categories
	Malagasy film categories
	Mauritanian film categories
	Mexican film categories

Table 25 (continued)

		Mongolian film categories
		Moroccan film categories
		Nepalese film categories
		New Zealand film categories
		Nicaraguan film categories
		North Korean film categories
		Norwegian film categories
		Pakistani film categories
		Palestinian film categories
		Peruvian film categories
		Philippine film categories
		Puerto Rican film categories
		Polish film categories
		Portuguese film categories
		Romanian film categories
		Russian film categories
		Senegalese film categories
		Serbian film categories
		Singaporean film categories
		Slovak film categories
		Slovenian film categories
		South African film categories
		South Korean film categories
		Spanish film categories
		Swedish film categories
		Swiss film categories
		Taiwanese film categories
		Tanzanian film categories
		Thai film categories
		Tunisian film categories
		Turkish film categories
		Vietnamese film categories
		Yugoslavian film categories
		Zimbabwean film categories
	Film categories by language	
	Film categories by genre	Action film categories
		Adventure film categories
		Animated film categories
		Art film categories
		Biographical film categories
		Children's film categories
		Comedy film categories
		Concert film categories
		Crime film categories
		Documentary film categories
		Drama film categories

Table 25 (continued)

		Epic film categories
		Erotic film categories
		Fantasy film categories
		Horror film categories
		Musical film categories
		Mystery film categories
		Political film categories
		Religious film categories
		Romance film categories
		Science fiction film categories
		Sports film categories
		Spy film categories
		Teen film categories
		Thriller film categories
		Tragedy film categories
		War film categories
		Western film categories
	Film categories by theme/subgenre	Action/adventure-related film categories
		Business-related film categories
		Comedy-related film categories
		Crime-related film categories
		Dance-related film categories
		Disaster/doomsday-related film categories
		Environment-related film categories
		Ethnicity/ethnography-related film categories
		Food-related film categories
		Friendship-related film categories
		History-related film categories
		Ideology/worldview-related film categories
		Legal system-related film categories
		Martial arts-related film categories
		Mathematics-related film categories
		Medical field-related film categories
		Monster/horror-related film categories
		Music-related film categories
		Propaganda-related film categories
		Racing/chase-related film categories
		Religion-related film categories
		Romance/sexuality-related film categories
		SciFi/fantasy-related film categories
		Sports-related film categories
		Travel/transportation-related film categories
		War-related film categories
		Film categories related to other themes/subgenres
	Film categories by subject	Film categories by generic subject matter
		Film categories by specific subject matter

Table 25 (continued)

		Documentary film categories by generic subject matter
		Documentary film categories by specific subject matter
Film categories by setting		Film categories by temporal setting
		Film categories by geographical setting
Film categories by plot source		Film categories by generic source
		Film categories by specific source
Film categories by shooting location		
Film categories by type		
Film categories by style		
Film categories by technology		
Film categories by crew/company		Film categories by director
		Film categories by producer
		Film categories by screenplay writer
		Film categories by music composer
		Film categories by actor/musician group
		Film categories by company
		Animated film categories by company
Film categories by award		Film categories by award winner
		Film categories by award-winning performance
Film categories by fictional character/milieu		Film categories by fictional character
		Film categories by fictional milieu
Film categories by specific series		
Film categories by specific film		
Film categories by topic/character		
Other film categories		
TV film/program categories		
Additional categories	Categories by topic/character portrayed in fiction/media/culture	Categories by topic/character portrayed in fiction
		Categories by topic/character portrayed in media
		Categories by topic/character portrayed in popular culture
Categories by fictional character/prop/setting		
Categories by non-film/TV medium/genre		Categories concerning book
		Categories concerning novel
		Categories concerning play
		Categories concerning book
		Categories concerning musical
		Categories concerning song/album
		Categories concerning video/computer game
	Categories concerning anime and manga	
More categories		

Out of the four Wikipedia page sections that were used for extracting information on films — the abstract section, the infobox section, the categories section, and the film cast section — the film cast section was by far the hardest to process. To begin with, at least over 60 variations have been observed with respect to how the film page section containing information on the cast members and/or their roles/characters may be differently titled or presented, as shown in Table 26. Beyond that, a seemingly infinite number of variations have been observed as to how (and how much of) the information concerning the cast members (with or without their role names) can be presented. Figures 36–41 show only a very few examples.

Table 26 PanAnthropon: Variations of film cast section titles considered

Cast and roles	Cast and Roles	Cast & Roles
Cast and characters	Cast and Characters	Cast & Characters
Cast and crew	Cast and Crew	Cast & Crew
Main cast	Main Cast	
Principal cast	Principal Cast	
Selected cast	Selected Cast	
Cast		
Casting		
Credits		
Crew and cast	Crew and Cast	Crew & Cast
Main crew and cast	Main Crew and Cast	Main Crew & Cast
Principal crew and cast	Principal Crew and Cast	Principal Crew & Cast
Selected crew and cast	Selected Crew and Cast	Selected Crew & Cast
Actors		
Main actors	Main Actors	
Principal actors	Principal Actors	
Selected actors	Selected Actors	
Roles		
Main roles	Main Roles	
Principal roles	Principal Roles	
Selected roles	Selected Roles	
Full cast	Full Cast	
Plot and cast	Plot and Cast	Plot & Cast
Featured cast	Featured Cast	
Voice cast	Voice Cast	
Voices		
Characters		
Main characters	Main Characters	
Principal characters	Principal Characters	
Selected characters	Selected Characters	

	Bette Davis as Julie Marsden		Henry Fonda as Preston Dillard
	George Brent as Buck Cantrell		Donald Crisp as Dr. Livingstone
	Fay Bainter as Aunt Belle Massey		

- Margaret Lindsay as Amy Bradford Dillard
- Richard Cromwell as Ted Dillard
- Henry O'Neill as General Theophilus Bogardus
- Spring Byington as Mrs. Kendrick
- John Litel as Jean La Cour
- Gordon Oliver as Dick Allen
- Janet Shaw as Molly Allen
- Theresa Harris as Zette
- Margaret Early as Stephanie Kendrick
- Irving Pichel as Huger
- Eddie Anderson as Gros Bat

Figure 36 Wikipedia: Sample film cast section #1

The 1953 *Desert Song* stars Kathryn Grayson as Margot and Gordon MacRae as the dashing outlaw leader (again called El Khobar). Latin tutor Paul Bonnard,^[1] Steve Cochran plays Claud Fontaine, El Khobar's rival for Margot's affections,^[2] and Raymond Massey plays Youseff, the villain.^[3] Comic relief is provided by Dick Wesson as Benny (not Benjy) Kidd,^[4] and Allyn McLerie is the voluptuous Az here depicted as Margot's father, rather than the hero's.^[6]

Figure 37 Wikipedia: Sample film cast section #2

Character	Species	Actor/Actress
Boog	Grizzly Bear (<i>Ursus arctos horribilis</i>)	Martin Lawrence
Elliot	Mule Deer (<i>Odocoileus hemionus</i>)	Ashton Kutcher
Beth	Human (<i>Homo sapiens</i>)	Debra Messing
Shaw	Human (<i>Homo sapiens</i>)	Gary Sinise
McSquizzzy	Eastern Gray Squirrel (<i>Sciurus carolinensis</i>)	Billy Connolly
Reilly	American Beaver (<i>Castor canadensis</i>)	Jon Favreau
Ian	Mule Deer (<i>Odocoileus hemionus</i>)	Patrick Warburton
Giselle	Mule Deer (<i>Odocoileus hemionus</i>)	Jane Krakowski
Mr. Weenie	Dachshund (<i>Canis lupus familiaris</i>)	Cody Cameron
Buddy	North American Porcupine (<i>Erethizon dorsatum</i>)	Matthew W. Taylor
Serge	Mallard (<i>Anas platyrhynchos</i>)	Danny Mann

Figure 38 Wikipedia: Sample film cast section #3

- **Phil Vischer** voices several roles:
 - **Bob the Tomato**
 - **Mr. Nezzar** as **Moses**
 - **Archibald Asparagus** as the **Commander of the Army of the Lord**
 - **Jimmy Gourd, Percy Pea, and Pa and Tom Grape** as **Israelites**
 - **Phillipe** as a defender of **Jericho**
- **Mike Nawrocki** voices several roles:
 - **Larry the Cucumber** as **Josh**
 - **Jerry Gourd** as an **Israelite**
 - **Jean Claude** as a defender of **Jericho**
- **Lisa Vischer** voices **Junior Asparagus** as a **Shepherd**
- **Jim Poole** voices **Scooter Carrot** as an **Israelite**

Figure 39 Wikipedia: Sample film cast section #4

- **George** (**Charles S. Dutton**) – the bus driver and trip organizer.
- **Jeremiah aka "Pop"** (**Ossie Davis**) – a downsized senior citizen who is an expert on African-American history.
- **Evan & Evan Jr. aka "Smooth"** (**Thomas Jefferson Byrd and De'aundre Bonds**) – an estranged father and son w together for 72 hours after Junior's arrest for petty theft.
- **Kyle & Randall** (**Isaiah Washington and Harry J. Lennix**) – a gay couple in the midst of breaking up.
- **Flip** (**Andre Braugher**) – a narcissistic actor.
- **Gary** (**Roger Guenveur Smith**) – a police officer who is half black and half white.
- **Xavier** (**Hill Harper**) – a UCLA Film School student who is making a documentary.
- **Jamal** (**Gabriel Casseus**) – a former gangster turned Muslim seeking redemption.
- **Jay** (**Bernie Mac**) – a bubble gum company owner.
- **Mike** (**Steve White**) – a conspiracy theorist who thinks the march is a plot to gather one million black men in one
- **Craig** (**Albert Hall**) – the original bus driver who is dealing with his teenage daughter's pregnancy.

Figure 40 Wikipedia: Sample film cast section #5

Segment 1 Honeymoon Suite	Segment 3 Room 309	Segment 2 Room 404	Segment 4 Penthouse
Tim Roth as Ted the Bellhop			
Valeria Golino as Athena	Antonio Banderas as Man	David Proval as Sigfried	Quentin Tarantino as Chester Rush
Madonna as Elspeth	Tamlyn Tomita as Wife	Jennifer Beals as Angela	
Alicia Witt as Kiva	Lana McKissack as Sarah	Paul Skemp as Real Theodore	Paul Calderon as Norman
Sammi Davis as Jezebel	Danny Verduzco as Juancho	Lawrence Bender as Long Hair Yuppie Scum	Bruce Willis as Leo (uncredited)
Lili Taylor as Raven	Patricia Vonne as Corpse	Quinn Thomas Hellerman as Baby Bellhop	Kimberly Blair as Hooker (uncredited)
Ione Skye as Eva	Salma Hayek as TV dancing girl		
Amanda de Cadenet as Diana			
Marisa Tomei as Margaret			
Kathy Griffin as Betty			
Julie McClean as Left Redhead			
Laura Rush as Right Redhead			

Figure 41 Wikipedia: Sample film cast section #6

Extracting the film cast information required accommodating all such variations as shown in Figures 36–41 (and infinitely more). For this reason, two programs were written in order to take care of the cases in which, for each entry, the actor name(s) appear(s) before the role name(s) and the cases in which the role name(s) appear(s) before the actor name(s), respectively, although there were cases where even such ordering was not followed consistently. The two programs used various pattern matching rules in order to identify and separate actor names, role names, role types (e.g., voice role, cameo role, guest role, etc.), role descriptions, and casting type notes (e.g., uncredited, unbilled, etc.). (In general, role description sentences/paragraphs were discarded, except for the short phrases that were incorporated into role names.)

The facts concerning a film and its cast members were entered into the database using the attribute `has_cast_member`. The information on the role(s) played by a cast member, if any, was stored in the `note` field (used to store context information for a given fact, as will be described in Section 6.2).

It may be mentioned that the role information entered into the `note` field was subject to a standardization process, before the `has_cast_member` facts were entered into the database, so that the role names and other details could be formatted as in a consistent manner as possible. In that way, at least some of the role names pertaining to well-known figures were resolved to the same formats (e.g., “King Richard I — Richard the Lionheart”) even though one and the same role name could appear in many variations (e.g., “King Richard I”, “Richard I of England”, “Richard the Lionheart”, “Richard the Lionhearted”, etc.).

However, no attempt was made to completely disambiguate or differentiate between identical role names that appear in multiple films. (For example, a role named “John” could appear in many

films, representing different characters.) In general, no attempt was made to differentiate between different characters, except some that represent well-known historic or fictional characters or recurring characters in the film series. The decision was based on the consideration that it could be both useful and interesting to be able to retrieve all films that share identical role names that may or may not represent identical characters.

Table 27 shows the types of film-oriented facts extracted through direct information extraction.

Table 27 PanAnthropon: Types of film-centric facts extracted

Entity	Attribute	Value	Note (Context)
film	type	work	
	subtype	film	
	associated_with_category	category	original category name
	also_known_as	[literal]	
	alt_name_url	[literal]	
	directed_by	director	director status
	produced_by	producer	producer status
		film company	
	written_by	writer	writing type / original source
	narrated_by	narrator	
	starring	actor	
	music_by	musician	musician role / music source
	cinematography_by	cinematographer	
	animation_by	animator	
	editing_by	editor	editor status
		film company	
	from_studio	film company	
	distributed_by	film company	regions of distribution / applicable years
		distributor	
	released_in_year	year	release date/year-month/year
	released_in_year-month	year-month	
	released_in_month	month	
	released_on_month-day	month-day	
	released_on_date	date	
	has_running_time	[literal]	varying running times
	from_country	country	constituent country
	in_language	language	
has_cast_member	actor	role name(s) / role type(s)	

(The entries given under “Entity” and “Value” represent the subtypes of corresponding entities, except for the attributes `also_known_as`, `alt_name_url`, and `has_running_time`, which take literals as values. The entries shown in bold (**work** and **film**) represent the corresponding values themselves for the given attributes, namely, `type` and `subtype`. The entries shown under “Attribute” are the actual attribute names themselves. The entries shown under “Note (Context)” represent the types of contextual information that may be optionally entered.)

In addition to the film pages, the pages on Academy Awards and Golden Globe Awards were processed in order to extract information on the award events and award winners/nominees. This was another time-consuming task due to the fact that most pages concerning particular awards presented the award/nominee record information differently than the others.

Information on the Academy Awards was extracted as follows: First, the “List of Academy Awards ceremonies” (http://en.wikipedia.org/wiki/List_of_Academy_Awards_ceremonies) page was processed to extract the event date and corresponding film year(s) for each ceremony from the 1st (1929) to the 82nd (2010). Each award ceremony was entered into the database as an entity of `type event`, `subtype film award event`. Information on the event date was also entered by using the attributes `held_in_year`, etc. Second, `Academy Award` was entered as an entity of `type cultural convention`, `subtype film award`. Each particular award (e.g., `Academy Award for Best Picture`) was also entered with the same entity type and subtype, and was then related to the `Academy Award` entity by using the attribute `belongs_to_award`. Finally, each award page was used to extract the winner/nominee facts for each ceremony. The extracted facts were entered into the database in terms of the relations between the award winner/nominee, on the one hand, and the award, the award event, the award event year, etc., the honored film year(s), and the film, on the other. A similar process was used for the Golden Globe Awards.

Table 28 shows the types of film-award-related facts extracted through the procedure as described above. In the table “person” stands for one of its subtypes: director, producer, or actor.

Table 28 PanAnthropon: Types of film-award-related facts extracted

Entity	Attribute	Value
film award event	type	event
	subtype	film award event
	held_for_year	year
	held_in_year	
	held_in_year-month	year-month
	held_in_month	month
	held_on_month-day	month-day
	held_on_date	date
	associated_with_category	category
film award	type	cultural convention
	subtype	film award
	belongs_to_award	film award
film	has_won_award	film award
	has_won_award_at_event	film award event
	has_won_award_for_year	year
	has_won_award_in_event_year	
	has_won_award_in_event_year-month	year-month
	has_won_award_in_event_month	month
	has_won_award_on_event_month-day	month-day
	has_won_award_on_event_date	date
	nominee_for_award	film award
	nominee_for_award_at_event	film award event
	nominee_for_award_for_year	year
	nominee_for_award_in_event_year	
	nominee_for_award_in_event_year-month	year-month
	nominee_for_award_in_event_month	month
	nominee_for_award_on_event_month-day	month-day
nominee_for_award_on_event_date	date	
person (director or producer or actor)	has_won_award	film award
	has_won_award_at_event	film award event
	has_won_award_for_film	film
	has_won_award_for_year	year
	has_won_award_in_event_year	
	has_won_award_in_event_year-month	year-month
	has_won_award_in_event_month	month
	has_won_award_on_event_month-day	month-day
	has_won_award_on_event_date	date
	nominee_for_award	film award
	nominee_for_award_at_event	film award event
	nominee_for_award_for_film	film
	nominee_for_award_for_year	year
	nominee_for_award_in_event_year	
	nominee_for_award_in_event_year-month	year-month
nominee_for_award_in_event_month	month	
nominee_for_award_on_event_month-day	month-day	
nominee_for_award_on_event_date	date	

6.1.3 Indirect Derivation of Information

The facts directly extracted by processing Wikipedia pages were themselves used to indirectly derive more classes, entities, attributes, and facts. The process involved: (1) deriving `starring` or `has_cast_member` facts in case only one type of facts could be extracted for a given film; (2) deriving inverse attributes/facts from the facts extracted from the infobox section; (3) deriving film role entities from `has_cast_member` facts, then deriving attributes/facts on those entities; (4) deriving entities, attributes, and facts concerning temporal/geographical inclusion relations; (5) deriving inverses of `associated_with_category` facts, deriving classes, entities, and facts by using super-categories and `associated_with_category` facts, and deriving inverse attributes/facts for the facts; (6) deriving indirect/inverse attributes and facts from film-award-related facts; and (7) deriving person name entities and related attributes and facts.

Some film pages have infoboxes that contain the starring information but do not have the film cast information section. The reverse is the case for some other film pages. The derivation of one type of information or the other in such cases was done by using the common-sense knowledge about the relation between starring actors and cast members of a film, namely, that all starring actors can be considered as cast members of the film but that not all cast members may be considered as starring actors (although some infoboxes tend to list (almost) all cast members as starring actors). The two opposite cases, i.e., starring-information/no-cast-information case and no-starring-information/cast-information case, were processed in turn for all films at once. First, in order to process the first case, all films that have at least one `starring` fact but no `has_cast_member` fact were identified by querying the database. For those films, a `has_cast_member` fact was entered for each `starring` fact by using the same value entity. Second, in order to process the second case, all films that have at least one `has_cast_member` fact but no `starring` fact were identified. In this case, only the first four `has_cast_member`

fact tuples were used to derive `starring` facts. The decision is based on the observation that usually those actors who have more important roles are listed in the film cast section before those who have relatively minor roles. The number four was chosen as a reasonable cutoff line.

Deriving inverse facts from the facts extracted from the infobox section were done for each attribute by transposing the entity in the subject position with the entity in the object position and by relating them with an inverse attribute (i.e., `<film directed_by director>` → `<director directed_film film>`, where `film` and `director` stand for a particular film entity and a particular director entity). The inverse facts were entered into the database by adding the film release year information in the `note` field so that such facts, when retrieved in response to queries, could be automatically sorted by the film release year. Table 29 shows the inverse attributes derived from the attributes extracted from the infobox section.

Table 29 PanAnthropon: Inverse attributes derived from infobox-based attributes

Extracted Fact Attribute	Inverse Fact Attribute
<code>directed_by</code>	<code>directed_film</code>
<code>produced_by</code>	<code>produced_film</code>
<code>written_by</code>	<code>provided_writing_for_film</code>
<code>narrated_by</code>	<code>narrated_film</code>
<code>starring</code>	<code>starred_in_film</code>
<code>music_by</code>	<code>provided_music_for_film</code>
<code>cinematography_by</code>	<code>provided_cinematography_for_film</code>
<code>animation_by</code>	<code>provided_animation_for_film</code>
<code>editing_by</code>	<code>edited_film</code>
<code>from_studio</code>	<code>studio_for_film</code>
<code>distributed_by</code>	<code>distributed_film</code>
<code>released_in_year</code>	<code>year_of_release_of_film</code>
<code>released_in_year-month</code>	<code>year-month_of_release_of_film</code>
<code>released_in_month</code>	<code>month_of_release_of_film</code>
<code>released_on_month-day</code>	<code>month-day_of_release_of_film</code>
<code>released_on_date</code>	<code>date_of_release_of_film</code>
<code>from_(or_co-produced_in)_country</code>	<code>county_of_film</code>
<code>in_language</code>	<code>language_of_film</code>

The `has_cast_member` facts derived from the film cast section were used first to derive inverse facts concerning a film and its cast members, and then to derive entities corresponding to the roles (characters) played by the cast members and also the facts involving those entities with respect to the films and with respect to the actors playing the roles.

First, `cast_member_of_film` facts were derived as inverses of `has_cast_member` facts. In addition, `has_cameo_appearance_by` and `has_guest_appearance_by` facts were derived depending on the role types, and `cameoed_in_film` and `guest_starred_in_film` facts were derived as inverses.

Next, the content of the `note` field for each `has_cast_member` fact was processed in order to derive film role entities. Each of the role names identified was entered into the database as an entity of type `concept`, subtype `film role`, after it was confirmed that the database did not contain that entity. (In deriving film role entities, same role names with different role types were considered as different entities. For example, the role name “John Smith” is considered as separate from “John Smith [voice]” and from “John Smith [cameo]”.) The names of role entities have “ (role)” appended at the end so that they can be distinguished from regular person entity names, e.g., `Sean Connery` vs. `Sean Connery (role)`.

Finally, with a total of 74,461 distinct film role entities derived through the procedure described above, facts concerning the relationships between those roles and the films in which they appear, and between the former and the actors who played them, were derived.

Table 30 shows the types of facts derived by using the facts extracted from the film cast section.

Table 30 PanAnthropon: Types of facts derived from film-cast-related facts

Entity	Attribute	Value	Note
actor	cast_member_of_film	film	film release year + role name
	cameoed_in_film		
	guest_starred_in_film		
	played_role	film role	film release year + film name
film	has_cameo_appearance_by	actor	film release year + role name
	has_guest_appearance_by		
	has_role	film role	film release year + actor name
film role	appears_in_film	film	film release year + actor name
	played_by	actor	film release year + film name

The derivation of entities, attributes, and facts concerning temporal/geographical inclusion relations was based on the common-sense knowledge concerning such relations. (Such facts were derived only for the temporal/geographical entities extracted.) Table 31 shows the types of temporal/geographical inclusion facts.

Table 31 PanAnthropon: Types of facts derived about temporal/geographical inclusion relations

Entity	Attribute	Value
continent	contains	country
country	contains	constituent country
		state/province/region
		city
	belongs_to	continent
constituent country	contains	state/province/region
		city
	belongs_to	country
state/province/region	contains	city
	belongs_to	constituent country
		country
city	belongs_to	state/province/region
		constituent country
		country
century	includes_decade	decade
decade	includes_year	year
	belongs_to_century	century
year	includes_year-month	year-month
	belongs_to_decade	decade
month	includes_month-day	month-day
month-day	belongs_to_month	month
year-month	includes_date	date
	belongs_to_year	year
date	belongs_to_year-month	year-month

The indirect information derivation process concerning categories involved: (i) deriving inverse facts for `associated_with_category` facts, (ii) deriving classes, entities, attributes, and facts by using the super-categories shown in Table 25 and the `associated_with_category` facts, and (iii) deriving inverse facts for the facts derived through (ii).

Derivation of `associated_with_film` facts as inverses of `associated_with_category` facts was done similarly as the derivation of other inverse facts. The difference, however, is that, unlike other inverse facts, the `associated_with_film` facts, which are category-centric facts having each category in the position of the subject, were stored in a separate table in the database, apart from the table that contains entity-centric facts. These category-centered facts are mainly used for the “Category-Based Entity Browsing” function of the interface.

As mentioned in Chapter 5, the process of constructing the film-domain-oriented ontology shown in Table 19 and the process of extracting/deriving information on the film domain in this research mutually informed each other. A prime example of this two-way process is the derivation of classes that represent film-related concepts and technologies and subsequently of entities that belong to those classes as well as facts involving those entities.

First, most of the leaf classes under `concept` and `technology` in the ontology in Table 19 were derived from some of the level-2 super-categories in Table 25, as shown in Table 32.

Next, by processing some of the `associated_with_category` facts, new entities were created and new facts were derived by using relevant attributes. (The facts derived involve three cases, depending on whether the value corresponds to a newly-created entity that belongs to one of the classes in Table 32, or a new entity of an existing class, or an existing entity of an existing class.)

Table 32 PanAnthropon: Classes derived from super-categories

Level-2 Super-Category	Class	Subclass
Film categories by genre	film genre	
Film categories by theme/subgenre	film subgenre	
Film categories by subject	film subject	generic subject
		specific subject
Film categories by setting	film setting	temporal setting
		geographical setting
Film categories by plot source	film plot source	generic subject
		specific subject
Film categories by type	film type	
Film categories by style	film style	
Film categories by technology	film technology	

For example, by using the `associated_with_category` facts involving (regular) categories that are classified under the (level-3) super-categories subsumed by the (level-2) super-category `Film categories by genre`, entities of type concept, subtype `film genre`, were derived from the categories that serve as the values for those facts, e.g., `Action thriller films` → `Action Thriller Film (genre)`, and then facts relating the films and the new entities were derived using `belongs_to_genre_of`, e.g., `<The Rock (film), belongs_to_genre_of, Action Thriller Film (genre)>`. Since `Action thriller films` is classified under `Action film categories`, an additional (transitive) fact, i.e., `<The Rock (film), belongs_to_genre_of, Action Film (genre)>`, was also derived. (Similarly, additional facts were derived from the facts concerning the temporal/geographical settings and shooting locations of films, based on the inclusion relations that exist between temporal/geographical units.) (In addition, the `associated_with_category` facts involving the categories classified under the super-categories `Film categories by country` and `Film categories by language` were used to derive `from_country` facts and `in_language` facts to supplement those extracted from the infoboxes.)

Finally, the inverse facts corresponding to the facts derived as above, e.g., `<Action Thriller Film (genre), genre_of_film, The Rock (film)>`, were also derived and stored.

Table 33 shows the types of facts derived by using categories and facts concerning categories.

Table 33 PanAnthropon: Types of facts derived from category-related facts

Entity	Attribute	Value	
film	belongs_to_genre_of	film genre	
	belongs_to_theme/subgenre_of	film subgenre	
	is_about_(generic_subject)	generic subject	
	is_about_(specific_subject)	specific subject	
	is_documentary_about_(generic_subject)	generic subject	
	is_documentary_about_(specific_subject)	specific subject	
	based_on_(generic_source)	generic source	
	based_on_(specific_source)	specific source	
	set_in_(temporal_setting)	century	
		decade	
		temporal setting	
	set_in_(geographical_setting)	continent	
		country	
		constituent country	
		state/province/region	
		city	
		geographical setting	
shot_in_(shooting_location)	continent		
	country		
	constituent country		
	state/province/region		
	city		
shot_in_(technology)	film technology		
belongs_to_film_type_of	film type		
belongs_to_film_style_of	film style		
film genre	genre_of_film	film	
film subgenre	theme/subgenre_of_film	film	
generic subject	generic_subject_of_film	film	
	generic_subject_of_documentary_film		
specific subject	specific_subject_of_film	film	
	specific_subject_of_documentary_film		
generic source	generic_source_for_film	film	
specific source	specific_source_for_film	film	
century	temporal_setting_for_film	film	
decade	temporal_setting_for_film	film	
temporal setting	temporal_setting_for_film	film	
continent	geographical_setting_for_film	film	
	shooting_location_for_film		
country	geographical_setting_for_film	film	
	shooting_location_for_film		
constituent country	geographical_setting_for_film	film	
	shooting_location_for_film		
state/province/region	geographical_setting_for_film	film	
	shooting_location_for_film		
city	geographical_setting_for_film	film	
	shooting_location_for_film		
geographical setting	geographical_setting_for_film	film	
film technology	technology_associated_with_film	film	
film type	type_of_film	film	
film style	style_of_film	film	

Film-award-related facts were also used to derive indirect/inverse attributes and facts, using similar procedures as those described above. Table 34 (continued on p.112) shows the types of facts thus derived.

Table 34 PanAnthropon: Types of facts derived from film-award-related facts

Entity	Attribute	Value
film award event	event_for_award_winning_by_film	film
	event_for_award_nomination_by_film	
	event_for_award_winning_by_person	director / producer / actor
	event_for_award_nomination_by_person	
film award	includes_award	film_award
	won_by_film	film
	nominated_for_by_film	
	won_by_person	director / producer / actor
	nominated_for_by_person	
	won_by_person_for_film	film
	nominated_for_by_person_for_film	
film	produced_winner_of_award	film award
	produced_nominee_for_award	
	produced_award_winner	director / producer / actor
	produced_award_nominee	
	produced_best-director-award-winning_director	director
	produced_best-director-award-nominated_director	
	produced_best-actor-award-winning_actor	actor
	produced_best-actor-award-nominated_actor	
	produced_best-actress-award-winning_actress	
	produced_best-actress-award-nominated_actress	
	produced_best-supporting-actor-award-winning_actor	
	produced_best-supporting-actor-award-nominated_actor	
	produced_best-supporting-actress-award-winning_actress	
produced_best-supporting-actress-award-nominated_actress		
director	has_won_best_director_award_for_film	film
	nominee_for_best_director_award_for_film	
actor	has_won_best_actor_award_for_film	film
	has_won_best_actress_award_for_film	
	has_won_best_supporting_actor_award_for_film	
	has_won_best_supporting_actress_award_for_film	
	nominee_for_best_actor_award_for_film	
	nominee_for_best_actress_award_for_film	
	nominee_for_best_supporting_actor_award_for_film	
nominee_for_best_supporting_actress_award_for_film		
year	year_of_award_event	film award event
	event_year_of_award_winning_by_film	film
	event_year_of_award_nomination_by_film	
	event_year_of_award_winning_by_person	director / producer / actor
	event_year_of_award_nomination_by_person	
	film_year_for_award_event	film award event
	film_year_for_award_winning_by_film	film
	film_year_for_award_nomination_by_film	
	film_year_for_award_winning_by_person	director / producer / actor
film_year_for_award_nomination_by_person		

Table 34 (continued)

year-month	year-month_of_award_event	film award event
	event_year-month_of_award_winning_by_film	film
	event_year-month_of_award_nomination_by_film	
	event_year-month_of_award_winning_by_person	director / producer / actor
	event_year-month_of_award_nomination_by_person	
month	month_of_award_event	film award event
	event_month_of_award_winning_by_film	film
	event_month_of_award_nomination_by_film	
	event_month_of_award_winning_by_person	director / producer / actor
	event_month_of_award_nomination_by_person	
month-day	month-day_of_award_event	film award event
	event_month-day_of_award_winning_by_film	film
	event_month-day_of_award_nomination_by_film	
	event_month-day_of_award_winning_by_person	director / producer / actor
	event_month-day_of_award_nomination_by_person	
date	date_of_award_event	film award event
	event_date_of_award_winning_by_film	film
	event_date_of_award_nomination_by_film	
	event_date_of_award_winning_by_person	director / producer / actor
	event_date_of_award_nomination_by_person	

Finally, classes concerning person names were introduced into the ontology, and all names of person-type entities were processed to extract entities of type `concept`, subtype `given name` or `family name`, and to derive person-name-related attributes and facts as shown in Table 35.

Table 35 PanAnthropon: Types of person-name-related facts derived

Entity	Attribute	Value
person (any)	has_given_name	given name
	has_family_name	family name
	has_given_name_initial	[literal]
	has_family_name_initial	
	has_initials	
given name	given_name_of	person (any)
family name	family_name_of	person (any)

Upon completion of the indirect information derivation process, all unique attributes were entered into a database table with information on the applicable types of entities, types of values, and types of value entities; subsequently, all entity-centric facts were converted to attribute-centric facts and saved in a table with information on the types/subtypes of entities and value entities.

6.2 Information Storage and Organization

The information extracted/derived from Wikipedia has been stored in a MySQL database. This research did not focus on the efficiency/economics of data storage (i.e., database normalization or optimization), but rather on the expediency of retrieval and ease of inspection.

Three data models were used to organize the data extracted/derived through this research:

- Hierarchical Tree Model
- Common Relational Model
- Entity–Attribute–Value Model

By “Hierarchical Tree Model” it is meant, in this research, a data model that represents the data in the form of `<child_node, parent_node>`. The model is used to represent the hierarchical relationships among the elements (or nodes) within an ontology or taxonomy. The Hierarchical Tree Model was used for tables to store the class hierarchy (`Class`) and the super-category hierarchy (`Category_Super`). By “Common Relational Model” it is meant, in this research, a data model that represents the data in the form of `<element, field_1, field_2, ..., field_n>`. The Common Relational Model was used for tables to store the data on the entities (`Entity`), regular categories (`Category`), attributes (`Attribute`), and film page sections (`Page`). By “Entity–Attribute–Value Model” it is meant, in this research, a data model that represents the data in the form of `<entity, attribute, value>`. The Entity–Attribute–Value Model was used for tables to store entity-centric facts (`Entity_Fact`), category-centric facts (`Category_Fact`), and attribute-centered facts (`Attribute_Fact`).

Figures 42–44 show the database table schemas used in this research, grouped by the underlying data models.

Class	Category_Super
+classID	+categoryD
+className	+categoryName
+superclassID	+supercatID
+superclassName	+supercatName
+	+

Figure 42 PanAnthropon: Database table schemas based on Hierarchical Tree Model

Entity	Page	Category	Attribute
+entityID	+entityID	+categoryID	+attributeID
+entityName	+entityName	+categoryName	+attributeName
+entityURL	+entityURL	+origcatName	+inverseAttName
+entityType	+abstract_content	+categoryURL	+entityTypes
+entitySubtype	+abstract_text	+supercatID	+valueTypes
+has_wikipage	+infobox_content	+supercatName	+valueEntityTypes
+has_abstract	+categories_content	+topcatID	+
+has_infobox	+cast_content	+topcatName	
+has_categories	+	+entityIDs	
+has_cast_section		+numEntities	
+has_image		+	
+has_akas			
+			

Figure 43 PanAnthropon: Database table schemas based on Common Relational Model

Entity_Fact	Category_Fact	Attribute_Fact
+entityFactID	+categoryFactID	+attributeFactID
+entityID	+categoryID	+entityFactID
+entityName	+categoryName	+entityFactID
+attribute	+attribute	+attributeID
+valueID	+valueID	+attributeName
+valueName	+valueName	+entityType
+valueURL	+valueURL	+entitySubtype
+note	+note	+entityID
+	+	+entityName
		+entityURL
		+valueType
		+valueEntityType
		+valueEntitySubtype
		+valueID
		+valueName
		+valueURL
		+note
		+

Figure 44 PanAnthropon: Database table schemas based on Entity-Attribute-Value Model

Table 36 shows the structure of the table `Class` that contains the class hierarchy. Table 37 describes the table `Page` that contains film page sections. Tables 38 and 39 present the tables that contain entities (`Entity`) and entity-centric facts (`Entity_Fact`), respectively.

Table 36 PanAnthropon: Structure of database table `Class`

Field Name	Value Description	Value Datatype
classID	ID of class	INT
className	name of class	VARCHAR
superclassID	ID of superclass of class	INT
superclassName	name of superclass of class	VARCHAR

Table 37 PanAnthropon: Structure of database table `Page`

Field Name	Value Description	Value Datatype	Ref Table
entityID	ID of entity	INT	Entity
entityName	name of entity	TEXT	
entityURL	URL of entity page	TEXT	
abstract_content	abstract section (in HTML) of entity page	TEXT	
abstract_text	abstract section (in text) of entity page	TEXT	
infobox_content	infobox section of entity page	TEXT	
categories_content	categories section of entity page	TEXT	
cast_content	cast section of entity page	TEXT	

Table 38 PanAnthropon: Structure of database table `Entity`

Field Name	Value Description	Value Datatype
entityID	ID of entity	INT
entityName	name of entity	TEXT
entityURL	URL of entity page	TEXT
entityType	class corresponding to type of entity	VARCHAR
entitySubtype	class corresponding to subtype of entity	VARCHAR
has_wikipedia	whether or not entity page exists	TINYINT
has_infobox	whether or not entity page has infobox	TINYINT
has_categories	whether or not entity page has categories section	TINYINT
has_cast_section	whether or not entity page has cast section	TINYINT
has_image	whether or not entity page (infobox) has entity image	TINYINT
has_akas	whether or not entity has alternative name	TINYINT

Table 39 PanAnthropon: Structure of database table Entity_Fact

Field Name	Value Description	Value Datatype	Ref Table
entityFactID	ID of entity-centered fact	INT	
entityID	ID of entity	INT	Entity
entityName	name of entity	TEXT	
attribute	attribute	VARCHAR	
valueID	ID of value (class/category/entity)	INT	
valueName	name of value (class/category/entity/literal)	TEXT	
valueURL	URL of value (category/entity)	TEXT	
note	contextual information	TEXT	

Tables 40–42 describe the tables that contain the super-category hierarchy (*Category_Super*), regular categories (*Category*), and category-centric facts (*Category_Fact*), respectively. Tables 43 and 44 depict the tables that contain attributes (*Attribute*) and attribute-centric facts (*Attribute_Fact*), respectively.

Table 40 PanAnthropon: Structure of database table Category_Super

Field Name	Value Description	Value Datatype
categoryID	ID of super-category	INT
categoryName	name of super-category	TEXT
supercategoryID	ID of supercategory of super-category	INT
supercategoryName	name of supercategory of super-category	TEXT

Table 41 PanAnthropon: Structure of database table Category

Field Name	Value Description	Value Datatype	Ref Table
categoryID	ID of category	INT	
categoryName	name of category	TEXT	
origcatName	original name of category, if different	TEXT	
categoryURL	URL of category page	TEXT	
supercatID	ID of supercategory of category	INT	Category_Super
supercatName	name of supercategory of category	TEXT	
topcatID	ID of level-1 supercategory of category	INT	Category_Super
topcatName	name of level-1 supercategory of category	TEXT	
entityIDs	IDs of entities associated with category	TEXT	
numEntities	number of entities associated with category	INT	

Table 42 PanAnthropon: Structure of database table Category_Fact

Field Name	Value Description	Value Datatype	Ref Table
categoryFactID	ID of category-centered fact	INT	
categoryID	ID of category	INT	Category
categoryName	name of category	TEXT	
attribute	attribute	VARCHAR	
valueID	ID of value entity	INT	Entity
valueName	name of value entity	TEXT	
valueURL	page URL of value entity	TEXT	
note	original category name, if different	TEXT	

Table 43 PanAnthropon: Structure of database table Attribute

Field Name	Value Description	Value Datatype
attributeID	ID of attribute	INT
attributeName	name of attribute	TEXT
inverseAttName	name of inverse attribute, if exists	TEXT
entityTypes	applicable types of entities	TEXT
valueTypes	types of values	TEXT
valueEntityTypes	applicable types of value entities, if non-literal values	TEXT

Table 44 PanAnthropon: Structure of database table Attribute_Fact

Field Name	Value Description	Value Datatype	Ref Table
attributeFactID	ID of attribute-centered fact	INT	
entityFactID	ID of corresponding entity-centered fact	INT	Entity_Fact
attributeID	ID of attribute	INT	Attribute
attributeName	name of attribute	TEXT	
entityType	type of entity	TEXT	
entitySubtype	subtype of entity	TEXT	
entityID	ID of entity	INT	Entity
entityName	name of entity	TEXT	
entityURL	URL of entity	TEXT	
valueType	type of value (class/category/entity/literal)	TEXT	
valueEntityType	type of value entity	TEXT	
valueEntitySubtype	subtype of value entity	TEXT	
valueID	ID of value (class/category/entity)	INT	
valueName	name of value (class/category/entity/literal)	TEXT	
valueURL	URL of value (category/entity)	TEXT	
note	contextual information	TEXT	

6.3 Information Extraction Statistics

Table 45 presents the extraction statistics in terms of the number of records per each database table. (Note: Attribute-centric facts are of the same number as entity-centric facts, since the former and the latter are the same facts, albeit represented differently.) Table 46 shows the number of entities per entity type.

Table 45 PanAnthropon: Overall information extraction statistics

Database Table	Record Type	Count
Class	class	72
Page	film entity page	10,640
Entity	entity	209,266
Entity_Fact	entity-centric fact	2,354,931
Category_Super	super-category	215
Category	category	5,229
Category_Fact	category-centric fact	91,335
Attribute	attribute	190
Attribute_Fact	attribute-centric fact	2,354,931

Table 46 PanAnthropon: Number of entities per entity type

Entity Type	Count
person	69,171
work	11,355
organization	1,975
place	254
time	8,279
event	149
cultural convention	25
cultural artifact	114
concept	117,925
technology	19

In concluding this chapter, it may be emphasized that this research obtained the dataset shown in Table 45 by effectively processing a relatively small number of Wikipedia pages, in contrast to the comparable projects reviewed in Chapter 4, which used much larger source datasets and yet produced much fewer entities and facts relative to the size of the information sources used.

CHAPTER 7: IMPLEMENTATION: INTERFACE CONSTRUCTION

The Web-based search interface was implemented by using HTML, JavaScript, and JSP (in connection with the back-end MySQL database) on the Tomcat server. The interface is publicly accessible at: <http://dlib.ischool.drexel.edu:8080/sofia/PA/>.

7.1 Interface Examples

Given that the application domain of this research is the film domain, it would be appropriate to consider the kinds of search functions provided by well-known film-related sites on the Web in comparison to the kinds of functions provided by the interface constructed from this research.

The Internet Movie Database (IMDb) (<http://www.imdb.com/>) is touted as the “biggest, best, most award-winning movie site on the planet”. Figure 45 shows a partial snapshot of the homepage of IMDb.

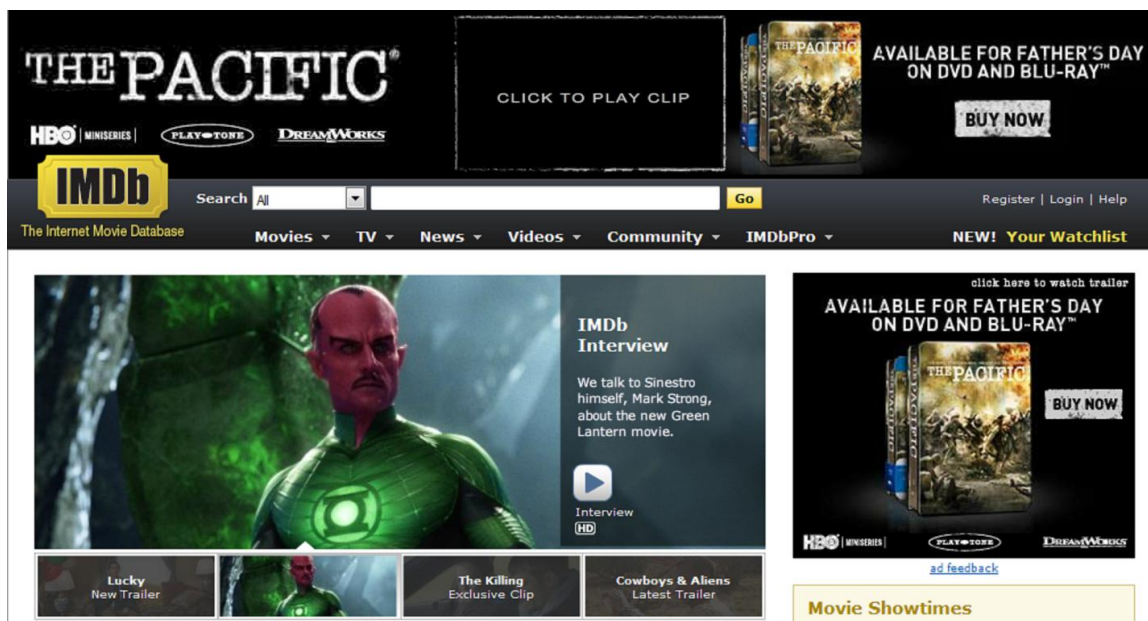


Figure 45 IMDb: Homepage

The “Movies” tab on the main menu bar on the homepage contains a dropdown menu shown in Figure 46. Depending on the menu selection, one is directed to a page containing the selected kind of information. For example, Figure 47 shows a partial snapshot of the “Now Playing” page.

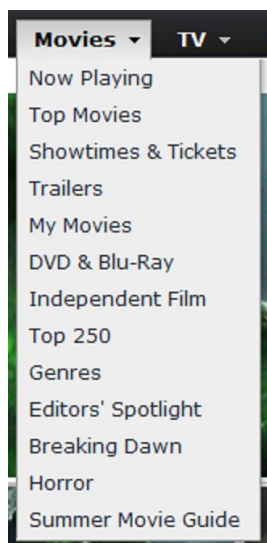



Figure 46 IMDb: Movie-related menus

Now Playing

U.S. Openings

JUNE 17th: Opening This Week | [Top 10](#) | [Jun](#) | [Jul](#) | [Aug](#) | [Sept](#) | [Oct](#) | [Nov](#) | [Dec](#) | [Jan](#)





Green Lantern


Director: [Martin Campbell](#)
Stars: [Ryan Reynolds](#), [Blake Lively](#), [Peter Sarsgaard](#) ([Full Cast](#))
Studio: Warner Bros. Pictures

The Plot: When he's granted a mystical green ring that bestows him with otherworldly powers, test pilot Hal Jordan (Reynolds) becomes the first human to earn membership into an intergalactic squadron tasked with keeping peace within the universe. His mission: to combat an enemy called Parallax, which threatens to destroy the universe's balance of power.

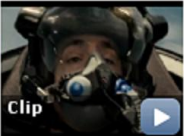
Photos ([See all 55](#) | [slideshow](#)) **Videos** ([see all 29](#))







Trailer
HD



Clip

Figure 47 IMDb: Now Playing page

More relevant to this research is the search function. Figure 48 shows the basic search menu on IMDb. As shown, the menu allows searching the content of the IMDb database by title, TV episode, (person) name, company name, etc. by entering keywords. For example, Figure 49 shows a partial snapshot of the search result page for a sample search by title (with keywords “The Lord of the Rings: The Return of the King”).

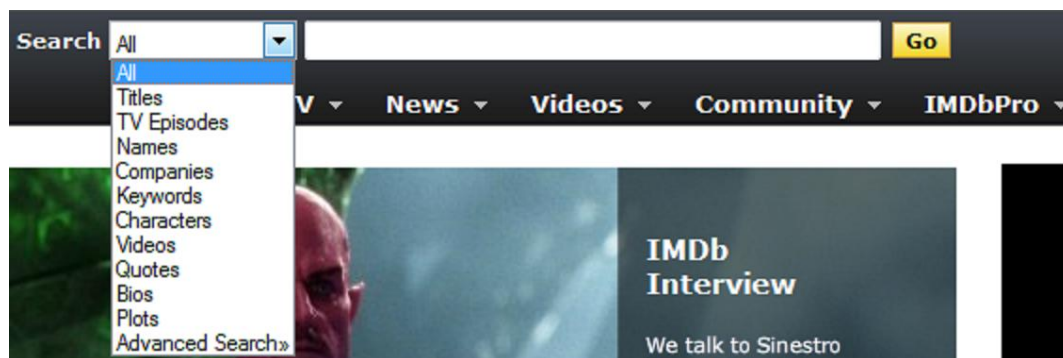


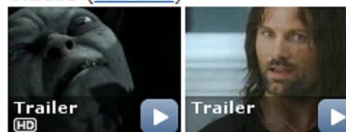
Figure 48 IMDb: Basic search menu

Media from [The Lord of the Rings: The Return of the King \(2003\)](#)

Photos ([See all 235](#) | [slideshow](#))



Videos ([see all 6](#))



Popular Titles (Displaying 1 Result)



1. [The Lord of the Rings: The Return of the King \(2003\)](#)
aka "The Return of the King" - USA (*short title*)

Titles (Exact Matches) (Displaying 1 Result)

1. [The Lord of the Rings: The Return of the King \(2003\) \(VG\)](#)

Titles (Partial Matches) (Displaying 1 Result)

1. [National Geographic: Beyond the Movie - The Lord of the Rings: Return of the King \(2003\) \(TV\)](#)

Titles (Approx Matches) (Displaying 17 Results)



1. [The Trouble of the Rings Returns: King-Size \(2004\) \(V\)](#)
2. [Lord of the Rings: Battle for Middle Earth II - Rise of the Witch King \(2006\) \(VG\)](#)
3. [The Lord of the Rings: The Fellowship of the Ring \(2001\)](#)
aka "The Fellowship of the Ring" - USA (*short title*)
⊞ aka "The Lord of the Rings: The Fellowship of the Ring: The Motion Picture" - USA (*promotional title*)

Figure 49 IMDb: Sample search-by-title result

The basic search menu shown in Figure 48 indicates an improved search function for the film domain — compared to the search functions provided by general-domain search engines such as Google — in that it makes it possible to specify the scope of search *within the given domain*. The ability to restrict the search as such is enabled by the incorporation of implicit domain knowledge in the menu options. That is to say, menu options such as “Titles”, “Names”, “Companies”, “Characters”, and “Plots” imply and exploit knowledge of the given domain, i.e., knowledge of the major types of entities/attributes that are relevant to the domain, even at a very basic level.

Even so, given the selection of the scope of search and given the input of query string, the retrieval of the results that match the given search criteria, when using the basic search function, is based on keyword matching. The fact that query–result matching is based on keyword matching is evidently illustrated in the sample query-by-title result in Figure 49, which shows that the results are sorted and presented according to the degree of keyword matching, i.e., exact matching, partial matching, and approximate matching.

The IMDb site, however, also offers an “Advanced Search” option, provided (rather confusingly and inconsistently) as one of the options in the basic search menu shown in Figure 48, along with other basic options. If one selects “Advanced Search” from the menu, one is then directed to the “Advanced Search” page (<http://www.imdb.com/search/>) shown partly in Figure 50. As shown, the types of advanced search that are available on IMDb include:

- Advanced Title Search
- Advanced Name Search
- Collaborations and Overlaps
- Title Text Search
- Name Text Search

Advanced Search

Welcome to Advanced Search, which gives you instant access to our complete catalog of 1,633,300 titles and 3,649,234 names.

Advanced Title Search
 Want to get a list of comedies from the 1970s that have at least 1000 votes and an average rating of 7.5 or higher? Use [Advanced Title Search](#).

Advanced Name Search
 Want a list of males in the database who are Virgos and over 6 feet tall? Use [Advanced Name Search](#).

Collaborations and Overlaps
 Want a list of titles in which both Brad Pitt and George Clooney appeared? Or a list of people who worked on both Forrest Gump and Apollo 13? Try searching [Collaborations and Overlaps](#).

Title Text Search
 Please select a section to search within (plot summary, quotes, trivia, etc) and enter a word to search for (e.g. "horses").

Plot

Name Text Search
 Please select a section to search within (mini biography, trivia, quotes) and enter a word to search for (e.g. "arrested").

Biographies

Figure 50 IMDb: Advanced Search page

As indicated in Figure 50, the Title Text Search and Name Text Search functions amount to extensions of the basic keyword search function. More relevant to this research are Advanced Title Search and Advanced Name Search, both of which are based on the specification of various conditions to be satisfied by the titles/names sought, as will be illustrated below. The advanced search for Collaborations and Overlaps is concerned with finding common cast/crew members, given two titles, or, conversely, finding common titles, given two cast/crew member names. It may be noted that a related, but more general, function for finding all commonalities between two entities of the same entity type and subtype is provided by the interface constructed from the this research. Furthermore, the interface also offers functions for finding both direct and indirect relations between two entities, regardless of their respective entity types and subtypes.

Figures 51–54 show the kinds of conditions that can be specified when using Advanced Title Search.

Title

e.g. The Godfather

Title Type

- Feature Film TV Movie TV Series TV Episode
 TV Special Mini-Series Documentary Video Game
 Short Film Video Unknown Work

Release Date [?](#)

 to

Format: YYYY-MM-DD, YYYY-MM, or YYYY

User Rating

 - to -

Number of Votes [?](#)

 to

Genres

- Action Adventure Animation Biography
 Comedy Crime Documentary Drama
 Family Fantasy Film-Noir Game-Show
 History Horror Music Musical
 Mystery News Reality-TV Romance
 Sci-Fi Sport Talk-Show Thriller
 War Western

Title Groups

- Oscar-Winning Best Picture-Winning Best Director-Winning
 Oscar-Nominated Emmy Award-Winning Emmy Award-Nominated
 Golden Globe-Winning Golden Globe-Nominated Razzie-Winning
 Razzie-Nominated National Film Board Preserved IMDb "Top 100"
 IMDb "Top 250" IMDb "Top 1000" IMDb "Bottom 100"
 IMDb "Bottom 250" IMDb "Bottom 1000" Now-Playing

Figure 51 IMDb: Part of Advanced Title Search interface #1

Title Data 

Alternate Versions	▲
Awards	☰
Blu-ray at Amazon.ca	
Blu-ray at Amazon.com	
Blu-ray at Amazon.de	
Blu-ray at Amazon.fr	
Blu-ray at Amazon.uk	
Book at Amazon.ca	
Book at Amazon.com	
Book at Amazon.de	▼

Companies

20th Century Fox
 Sony
 DreamWorks
 MGM
 Paramount
 Universal
 Walt Disney
 Warner Bros.

US Box Office Gross

to

US Certificates

G
 PG
 PG-13
 R
 NC-17

Color Info

Color
 Black & White
 Colorized

Countries

... Common Countries ...	▲
Argentina	☰
Australia	
Austria	
Belgium	
Brazil	
Bulgaria	▼

Keywords 

Search for a notable object, concept, style or aspect.

Languages

... Common Languages ...	▲
Arabic	☰
Bulgarian	
Chinese	
Croatian	
Dutch	
English	▼

Figure 52 IMDb: Part of Advanced Title Search interface #2

Filming Locations

MOVIEmeter

 to

Plot [?](#)

Search for words that might appear in the plot summary.

Production Status [?](#)

- | | | | |
|--|---|--|---|
| <input type="checkbox"/> Released | <input type="checkbox"/> Post-production | <input type="checkbox"/> Filming | <input type="checkbox"/> Pre-production |
| <input type="checkbox"/> Completed | <input type="checkbox"/> Script | <input type="checkbox"/> Optioned Property | <input type="checkbox"/> Announced |
| <input type="checkbox"/> Treatment/outline | <input type="checkbox"/> Pitch | <input type="checkbox"/> Turnaround | <input type="checkbox"/> Abandoned |
| <input type="checkbox"/> Delayed | <input type="checkbox"/> Indefinitely Delayed | <input type="checkbox"/> Active | <input type="checkbox"/> Unknown |

Cast/Crew

Runtime

 to

Sound Mix

- | | | |
|---|--|---|
| <input type="checkbox"/> Mono | <input type="checkbox"/> Silent | <input type="checkbox"/> Stereo |
| <input type="checkbox"/> Dolby Digital | <input type="checkbox"/> Dolby | <input type="checkbox"/> Dolby SR |
| <input type="checkbox"/> DTS | <input type="checkbox"/> SDDS | <input type="checkbox"/> Ultra Stereo |
| <input type="checkbox"/> 4-Track Stereo | <input type="checkbox"/> 70 mm 6-Track | <input type="checkbox"/> Vitaphone |
| <input type="checkbox"/> Dolby Digital EX | <input type="checkbox"/> De Forest Phonofilm | <input type="checkbox"/> DTS-Stereo |
| <input type="checkbox"/> Chronophone | <input type="checkbox"/> 6-Track Stereo | <input type="checkbox"/> DTS-ES |
| <input type="checkbox"/> Perspecta Stereo | <input type="checkbox"/> Cinephone | <input type="checkbox"/> 3 Channel Stereo |
| <input type="checkbox"/> Cinematophone | <input type="checkbox"/> Sonics-DDP | <input type="checkbox"/> 12-Track Digital Sound |
| <input type="checkbox"/> DTS 70 mm | <input type="checkbox"/> IMAX 6-Track | <input type="checkbox"/> Matrix Surround |
| <input type="checkbox"/> Sonix | <input type="checkbox"/> Sensurround | <input type="checkbox"/> Cinerama 7-Track |
| <input type="checkbox"/> Kinoplasticon | <input type="checkbox"/> Digitrac Digital Audio System | <input type="checkbox"/> Cinesound |
| <input type="checkbox"/> Phono-Kinema | <input type="checkbox"/> CDS | |

Figure 53 IMDb: Part of Advanced Title Search interface #3

Your Ratings

You must [login](#) or [register](#) to use this feature.

- Include All Titles
- Exclude Titles I've Seen
- Restrict to Titles I've Seen

Lists

You must [login](#) or [register](#) to use this feature.

MyMovies

You must [login](#) or [register](#) to use this feature.

Display Options

Display: sorted by



Figure 54 IMDb: Part of Advanced Title Search interface #4

All the menu groups and input fields in Figures 51–54 are optional, although at least one menu option must be selected or at least one input value must be entered in order for the search to work. Fit for the biggest movie site, the menu groups and input fields represent a variety of types of movie-related information. The menu groups and input fields roughly correspond to attributes, and the menu options and input values correspond to attribute values. Not all menu groups and input fields shown in Figures 51–54 are represented in the PanAnthropon FilmWorld interface, since not all of them are relevant or applicable in the latter. On the other hand, the latter also contains menu options (in the form of <attribute, value> pairs) not available on IMDb.

The result of a sample Advanced Title Search — with (the conjunction of) the following criteria: Title Type = Feature Film (AND) Release Date = 1970 to 1990 (AND) Genres = War (AND) Title Groups = Oscar-Winning (AND) Golden Globe-Winning — are partially shown in Figure 55. (It may be noted that the search results in Figure 55, which represent the films that exactly match the specified criteria, are sorted, rather than ranked, which is also the case for the search results returned by the PanAnthropon interface.)

Sort by: [MOVIEmeter](#) | [A-Z](#) | [User Rating](#) ▼ | [Num Votes](#) | [US Box Office](#) | [Runtime](#) | [Year](#) | [US Release Date](#)

1.		Apocalypse Now (1979) ★★★★★☆☆☆☆ 8.6/10 During the on-going Vietnam War, Captain Willard is sent on a dangerous mission into Cambodia to assassinate a renegade Green Beret who has set himself up as a God among a local tribe. Dir: Francis Coppola With: Martin Sheen , Marlon Brando , Robert Duvall Drama War 153 mins. 	Add to Watchlist
2.		The Deer Hunter (1978) ★★★★★☆☆☆☆ 8.2/10 An in-depth examination of the way that the Vietnam war affects the lives of people in a small industrial town in the USA. Dir: Michael Cimino With: Robert De Niro , Christopher Walken , John Cazale Drama War 182 mins. 	Add to Watchlist
3.		Platoon (1986) ★★★★★☆☆☆☆ 8.2/10 A young recruit in Vietnam faces a moral crisis when confronted with the horrors of war and the duality of man. Dir: Oliver Stone With: Charlie Sheen , Tom Berenger , Willem Dafoe Action Drama War 120 mins. 	Add to Watchlist
4.		Patton (1970) ★★★★★☆☆☆☆ 8.1/10 The World War II phase of the controversial American general's career is depicted. Dir: Franklin J. Schaffner With: George C. Scott , Karl Malden , Stephen Young Biography Drama War 172 mins. United States-GP	Add to Watchlist

Figure 55 IMDb: Sample Advanced Title Search result

The interface for Advanced Name Search is similar to the one for Advanced Title Search, although it contains fewer menu groups and input fields (Name, Birth Date, Star Sign, Birth Place, Death Date, Death Place, Gender, Height, Filmography, Biographies, and Lists).

In addition to (the links to) the five types of advanced search functions, the Advanced Search page on IMDb also provides a side menu, shown in Figure 56, which contains options for browsing titles and people. As shown, titles can be browsed by genre, country, language, year, and keyword. People can be browsed by gender and star sign. Figure 57 shows the page for browsing titles by genre. Pages for other browsing options are similar, except the page for browsing titles by keyword, which provides a function for searching titles by using a Movie Keyword Analyzer (MoKA) and a function for browsing keywords by letter, length, and count.



Figure 56 IMDb: Options for browsing titles/people

Genres

Love Westerns? Comedies? Film noir? Click on the genre links below to go to the biggest and best collection of films sorted by their type.

Action	Adventure	Animation	Biography
Comedy	Crime	Documentary	Drama
Family	Fantasy	Film-Noir	Game-Show
History	Horror	Music	Musical
Mystery	News	Reality-TV	Romance
Sci-Fi	Sport	Talk-Show	Thriller
War	Western		

Figure 57 IMDb: Interface for browsing titles by genre

FMDb (<http://fmdb.freebaseapps.com/>), an application developed by volunteer contributors, provides a “simple interface” to access the movie data on Freebase (<http://www.freebase.com/>). As shown in Figure 58, the interface is indeed simple. The single input field for keyword-based search provides suggestions, as shown in Figure 59. Interestingly, the suggestion list includes film/person names that do not contain (the characters in) the keyword (“Brav”) entered. However, no hint is provided as to how those films/people may be related to the keyword. Figure 60 shows the result of a sample search, which merely consists of a list of films whose titles contain the keyword entered. The FMDb interface does not provide any advanced search function beyond the keyword-based search function already available on the Freebase Web site.

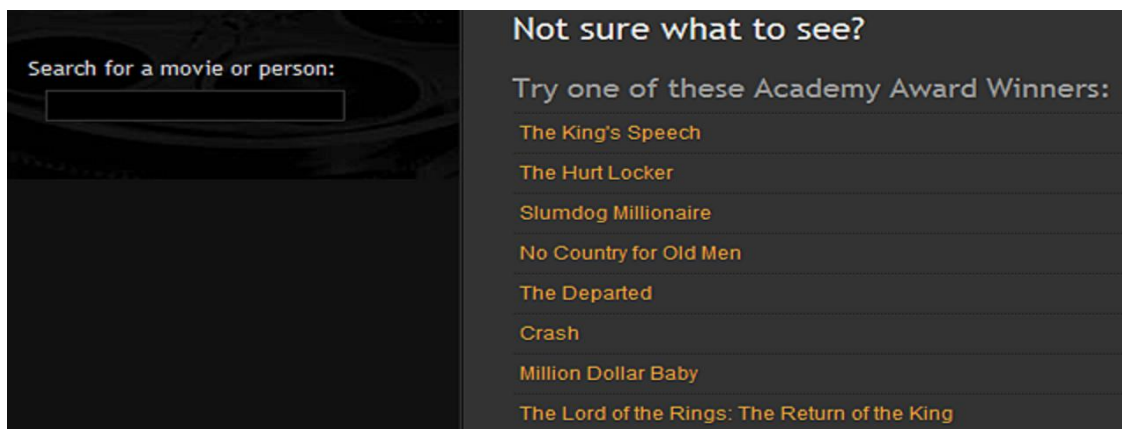


Figure 58 FMDb: Homepage

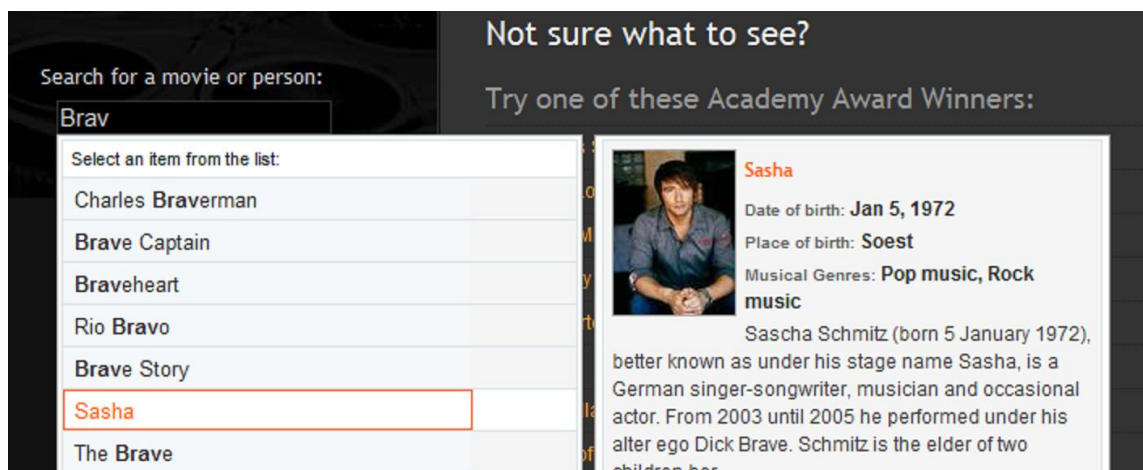


Figure 59 FMDb: Search suggestion menu

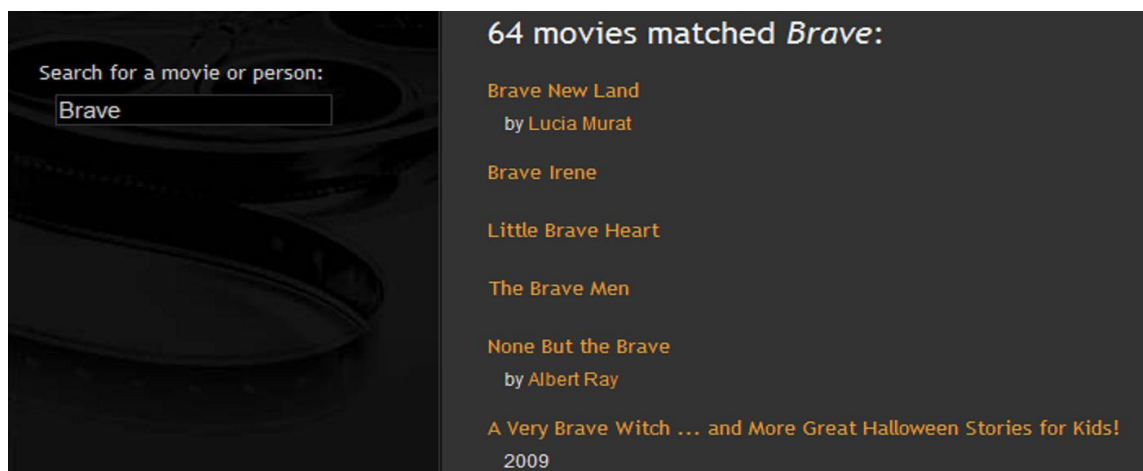


Figure 60 FMDb: Sample search result

The Web site of the LinkedMDB (<http://www.linkedmdb.org/>) project [Has09], reviewed in Chapter 4, only provides a browsing function. Figure 61 shows the “Start Exploring” section of the LinkedMDB homepage, which contains links for browsing film-related entities by type. If one clicks on the link “film”, one is directed to the page shown in Figure 62, which lists (some of) the films in the database (sorted in the ascending file-name-sorting order of the URI strings). If one clicks on any film title on the list, one is directed to a page that presents the film information in RDF format, similar to the one shown in Figure 15 (p.58).

Start Exploring

Browse linked open data for the following entities, using your web browser:
[actor](#) [cinematographer](#) [content_rating](#) [content_rating_system](#) [country](#) [director](#)
[dubbing_performance](#) [editor](#) [film](#) [film_art](#) [director](#) [film_awards_ceremony](#)
[film_casting_director](#) [film_character](#) [film_collection](#) [film_company](#)
[film_costume_designer](#) [film_crew](#) [gig](#) [film_crewmember](#) [film_critic](#) [film_cut](#)
[film_distribution_medium](#) [film_distributor](#) [film_featured_song](#) [film_festival](#)
[film_festival_event](#) [film_festival_focus](#) [film_festival_sponsor](#) [film_festival_sponsorship](#)
[film_film_company_relationship](#) [film_film_distributor_relationship](#) [film_format](#) [film_genre](#)
[film_job](#) [film_location](#) [film_production_designer](#) [film_regional_release_date](#)
[film_screening_venue](#) [film_series](#) [film_set_designer](#) [film_story_contributor](#) [film_subject](#)
[film_theorist](#) [interlink](#) [linkage](#) [run_music_contributor](#) [performance](#)
[personal_film_appearance](#) [personal_film_appearance_type](#) [producer](#)
[production_company](#) [special_film_performance_type](#) [writer](#)

Figure 61 LinkedMDB: Homepage

Example film

Home | Example data: [actor](#) [cinematographer](#) [content_rating](#) [content_rating_system](#) [country](#) [director](#) [dubbing_performance](#) [editor](#) [film](#) [film_art](#) [director](#) [film_awards_ceremony](#) [film_casting_director](#) [film_character](#) [film_collection](#) [film_company](#) [film_costume_designer](#) [film_crew](#) [gig](#) [film_crewmember](#) [film_critic](#) [film_cut](#) [film_distribution_medium](#) [film_distributor](#) [film_featured_song](#) [film_festival](#) [film_festival_event](#) [film_festival_focus](#) [film_festival_sponsor](#) [film_festival_sponsorship](#) [film_film_company_relationship](#) [film_film_distributor_relationship](#) [film_format](#) [film_genre](#) [film_job](#) [film_location](#) [film_production_designer](#) [film_regional_release_date](#) [film_screening_venue](#) [film_series](#) [film_set_designer](#) [film_story_contributor](#) [film_subject](#) [film_theorist](#) [interlink](#) [linkage](#) [run_music_contributor](#) [performance](#) [personal_film_appearance](#) [personal_film_appearance_type](#) [producer](#) [production_company](#) [special_film_performance_type](#) [writer](#)

- [Buffy the Vampire Slayer](#)
http://data.linkedmdb.org/resource/film/1
- [Final Fantasy: The Spirits Within](#)
http://data.linkedmdb.org/resource/film/10
- [Disraeli](#)
http://data.linkedmdb.org/resource/film/100
- [Akagi](#)
http://data.linkedmdb.org/resource/film/1000
- [Ä Ä Ä HOLIC](#)
http://data.linkedmdb.org/resource/film/1002
- [TaiyÄ o Nusunda Otoko](#)
http://data.linkedmdb.org/resource/film/1003
- [Chance](#)
http://data.linkedmdb.org/resource/film/1004

Figure 62 LinkedMDB: Interface for browsing films

Netflix (<http://www.netflix.com/>) is one of the largest DVD-rental sites on the Web. Figure 63 shows the main menu bar on the Netflix homepage. As shown, film and TV program titles can be browsed by genre, as on IMDb, as well as by a few other criteria. The main menu bar also contains a keyword search field. Figure 64 shows a search suggestion menu that appears upon keyword input. Perhaps a characteristic feature of the Netflix interface consists in the recommendation function illustrated in Figure 65, which is based on the ratings provided by the users. A similar function could be incorporated in the PanAnthropon interface by using a variety of semantic criteria. However, such a task is left as a potential future work.

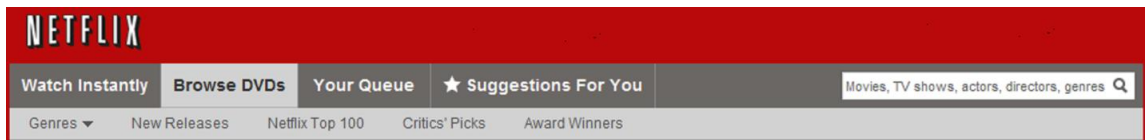


Figure 63 Netflix: Main menu bar

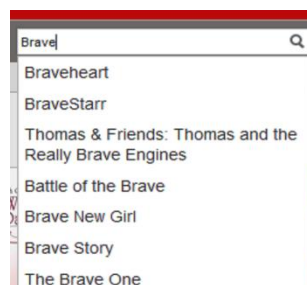


Figure 64 Netflix: Search suggestion menu



Figure 65 Netflix: Recommendation functionality

7.2 Interface Design and Implementation

7.2.1 Interface Functions

Six types of semantic search functions are currently provided by the PanAnthropon interface:

- (1) General Entity Retrieval Query (GERQ) function
- (2) Specific Entity-Centered Query (SECQ) function
- (3) Entity Commonality Finder Query (ECFQ) function
- (4) Direct Relation Finder Query (DRFQ) function
- (5) Indirect Relation Finder Query (IRFQ) function
- (6) Category-Based Entity Browsing (CBEB) function

The GERQ function is one that corresponds to the main research problem, namely, to demonstrate the capability of retrieving entities (and related facts) that directly match a given query. The “general” in the name indicates that this type of query searches for all entities that match the entity type, subtype, and conditions (i.e., `<attribute, value>` pairs) given in the query. This search function corresponds to the main Entity Ranking task in the INEX XER Track.

The SECQ function, in contrast, refers to the capability of retrieving all entity-centric facts (e.g., all facts centering on the film *The Lord of the Rings: The Return of the King*, i.e., all facts that have `The Lord of the Rings: The Return of the King` in the position of entity), given the type, subtype, and name of a specific entity. The function can thus be alternatively named Specific Entity Fact Query (SEFQ) or Entity Fact Retrieval Query (EFRQ).

The ECFQ function refers to retrieving commonalities between two specified entities that are of the same entity type and subtype (e.g., between two films, between two directors, etc.). Here commonalities mean common `<attribute, value>` pairs shared by the two entities.

The DRFQ function allows retrieving direct relations that hold between two specified entities (more precisely, direct relations that connect entity 1 to entity 2), regardless of their respective entity types and subtypes. In other words, the results of a query using this function consist of `<entity_1, relation, entity_2, note>` tuples, i.e., entity-centric fact tuples where entity 1 occupies the `entity` position, entity 2 occupies the `value` position, and `attribute` represents the relation between the two entities.

The IRFQ function allows retrieving 1-degree indirect relations between two specified entities, regardless of their respective types and subtypes. The results of a query using this function consist of `<entity_1, relation_e1-e3, entity_3, relation_e3-e2, entity_2>` tuples, where `entity_1` and `entity_2` stand for the specified two entities, `entity_3` stands for a third, intermediary entity, and `relation_e1-e3` and `relation_e3-e2` stand for the relation between entity 1 and entity 3 and the relation between entity 3 and entity 2, respectively. (Naturally, the actual entity that occupies the `entity_3` position in each tuple can be distinct.)

Finally, the CBEB function refers to retrieving (only) film entities by using the super-categories and categories assigned/associated to/with those entities.

Only the GERQ function is essential to fulfilling the purpose of this research. The SECQ function is used by itself and for presenting facts concerning the entities returned by other functions. The CBEB function demonstrates an enhanced browsing capability by providing much more specific criteria than are available on the other film-related Web sites.

In addition to the six main functions above, the interface also has a Slide function, which provides an image (if available) and brief introductory information for each film in the knowledge base.

7.2.2 Search/Retrieval Process

Figure 66 shows a flowchart representing the search/retrieval process using the PanAnthropon interface. (In the diagram, double-arrow connectors (\updownarrow) represent the actions that may be repeated. Albeit not explicitly represented in the diagram, one can always return to the start point at the end of (or during) the process.)

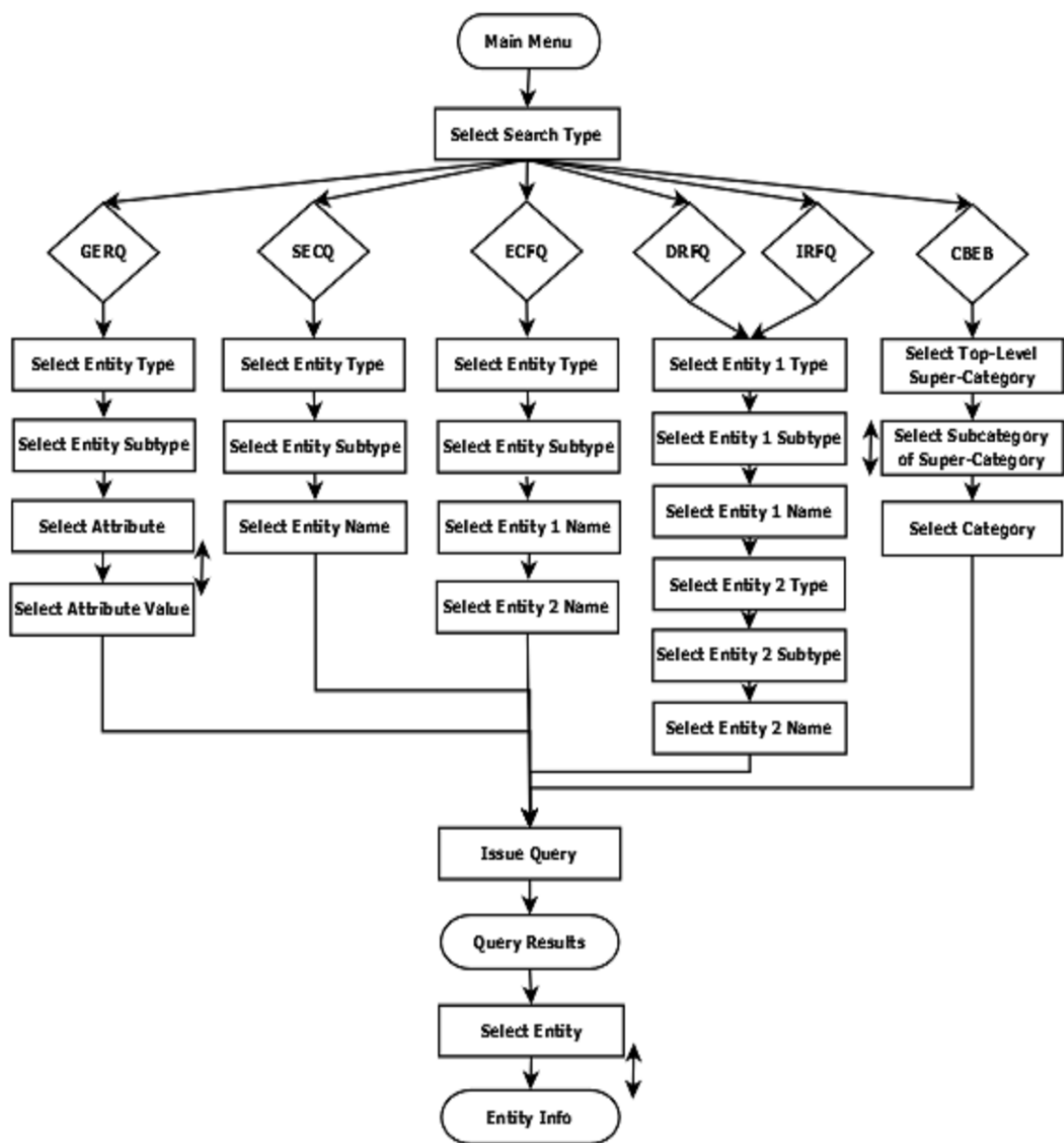


Figure 66 PanAnthropon: Flowchart of search process using interface

7.2.3 Interface Implemented

Figure 67 shows a snapshot of the PanAnthropon FilmWorld project homepage. As shown, the main menu bar contains links to the interfaces for various search functions as well to the pages containing the summary of research, the list of related publications, and the instructions for users.

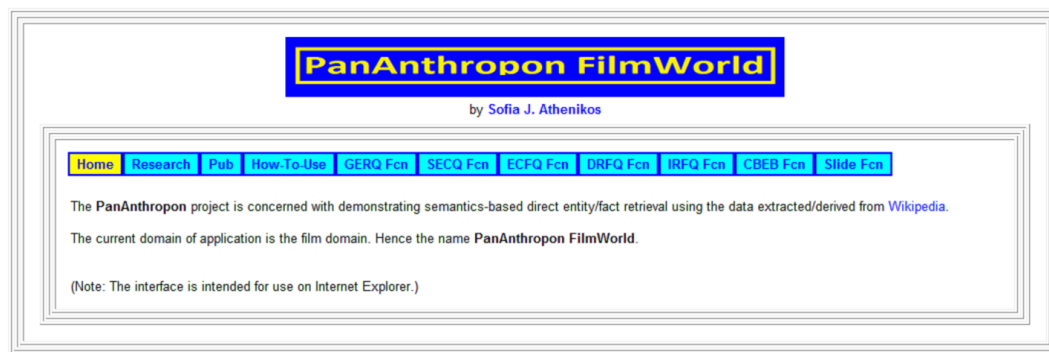


Figure 67 PanAnthropon: Homepage

The interface for each search/retrieval function is designed and implemented to be highly interactive and intelligent. (Thanks to the use of AJAX technology, implemented with JavaScript and JSP, all the user activities are processed and all the results are returned on a given page, without the need of directing the user to multiple pages, once a user selects a search option.) The interactive/intelligent interface works as follows: For GERQ: Given the user selection of an entity type, only relevant entity subtypes are displayed. In turn, given the selection of an entity subtype, only relevant attributes are presented. Finally, given the selection of an attribute, only relevant values are provided for user selection. (In case a large number of values correspond to an attribute, an input box is first presented so that the user can start typing in order to get suggestions for relevant values.) Similarly, for SECQ: Given the user selection of an entity type and an entity subtype, only relevant names are presented for user selection. Again, similarly, for CBEB: Given the selection of a top-level super-category, only relevant sub-categories are progressively presented until a leaf-level category is selected by the user. And similarly for other functions.

Figure 68 shows the initial screen of the GERQ interface page, which appears when the user clicks on **GERQ Fcn** on the main menu bar. (The user can click on **GERQ Fcn** anytime the user wishes to start a new search.)



Figure 68 PanAnthropon: GERQ function interface initial screen

Once the user clicks on **Want To Enter A Query?** button, an initial query form appears, which, at this stage, contains only a menu for entity type selection, as shown in Figure 69. The menu contains options representing different entity types, as shown in Figure 70.



Figure 69 PanAnthropon: GERQ function initial query form

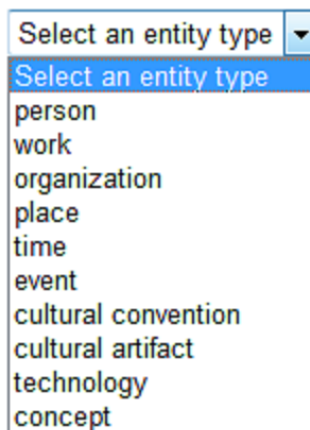


Figure 70 PanAnthropon: GERQ function entity type selection menu

Once the user selects an entity type, a new menu for entity subtype selection appears, as shown in Figure 71. The menu contains only entity subtypes that are relevant to the selected entity type, as shown in Figure 72. As indicated in Chapter 5, the interface presents a simplified classification of entity types and subtypes, which does not specify subtypes for entity type `person` and for entity type `organization`. (The reason for this decision is that most attributes relevant to entity type `person` are commonly applicable among its subtypes and similarly so for entity type `organization`.) Therefore, if the user selects either `person` or `organization` as entity type, then a simplified entity subtype menu appears, as shown in Figure 73.

Figure 71 PanAnthropon: GERQ function query form after entity type selection

Figure 72 PanAnthropon: GERQ function entity subtype selection menu

Figure 73 PanAnthropon: GERQ function simplified entity subtype selection menu

Once the user selects an entity subtype, a menu for attribute selection appears, as shown in Figure 74. (The "AND" in front of the attribute menu indicates that the entities to be retrieved must be of the selected entity type and entity subtype AND satisfy the condition represented by the attribute–value pair.) The menu for attribute selection contains only relevant attributes to choose from, according to the entity type and entity subtype selected, as shown in Figure 75.

person any (director, actor, producer, etc.)

AND Select an attribute

Figure 74 PanAnthropon: GERQ function query form after entity subtype selection

Select an attribute

- also_known_as
- alt_name_url
- cameoed_in_film
- cast_member_of_film
- directed_film
- distributed_film
- edited_film
- guest_starred_in_film
- has_given_name
- has_given_name_initial
- has_initials
- has_family_name
- has_family_name_initial
- has_won_award
- has_won_award_at_event
- has_won_award_for_film
- has_won_award_for_year_(yyyy)
- has_won_award_in_event_month_(mm)
- has_won_award_in_event_year_(yyyy)
- has_won_award_in_event_year-month_(yyyy-mm)
- has_won_award_on_event_date_(yyyy-mm-dd)
- has_won_award_on_event_month-day_(mm-dd)
- has_won_best_actor_award_for_film
- has_won_best_actress_award_for_film
- has_won_best_director_award_for_film
- has_won_best_supporting_actor_award_for_film
- has_won_best_supporting_actress_award_for_film
- narrated_film
- nominee_for_award

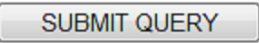
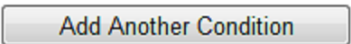

Figure 75 PanAnthropon: GERQ function attribute selection menu

Once the user selects an attribute, an input box may appear, as shown in Figure 76. Once the user starts typing in the input text box, a menu for value selection appears, as shown in Figure 77. The menu contains only those values that start with the letters the user entered in the input box, out of only those values that are relevant for the given attribute, as shown in Figure 78.

Figure 76 PanAnthropon: GERQ function query form after attribute selection #1

Figure 77 PanAnthropon: GERQ function query form after value text input

Figure 78 PanAnthropon: GERQ function value selection menu

In case there are a relatively small number of values to choose from, a menu for value selection immediately appears after attribute selection, as shown in Figure 79. Once the user selects a value for the selected attribute, buttons appear at the bottom of the query form, as shown in Figure 80. The user can submit the query by clicking on  button. Or, the user can add another condition by clicking on  button. Or, the user can remove the last condition by clicking on  button. Once the user submits the query, query processing starts, as shown in Figure 81.




Figure 79 PanAnthropon: GERQ function query form after attribute selection #2

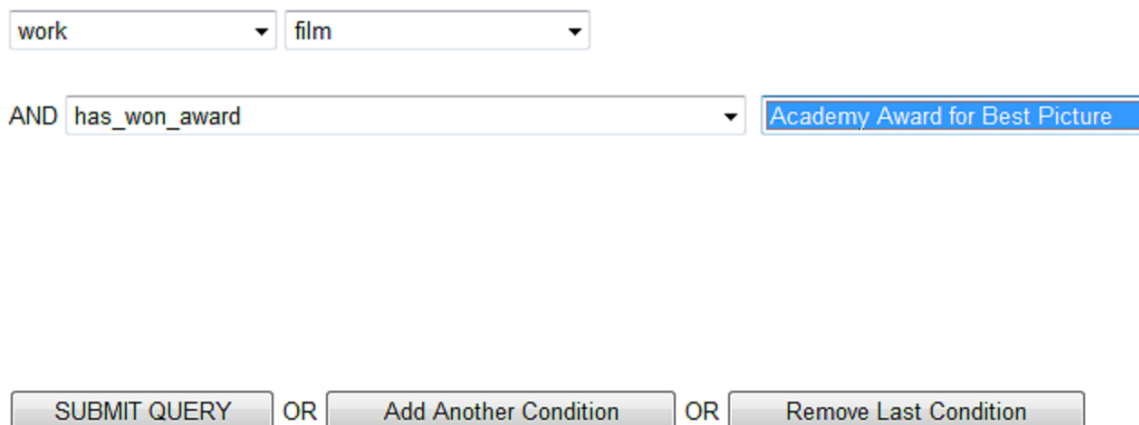


Figure 80 PanAnthropon: GERQ function query form after value selection

Query: Entity Type: **work**; Entity Subtype: **film**; Condition 1: [**has_won_award** -- **Academy Award for Best Picture**]

Processing query ... It may take a while ...


Figure 81 PanAnthropon: GERQ function query processing

Figure 82 shows a partial snapshot of the result of a sample GERQ query, presented in the alphabetical order of entity names. As shown, the result does not consist of a simple list of entity names, but it provides query-relevant fact(s) concerning each entity in the form of <entity, attribute, value, note>. (In the case of film entities, thumbnail images and release years are also presented, as shown.)


Query: Entity Type: work; Entity Subtype: film; Condition 1: [has_won_award -- Academy Award for Best Picture]

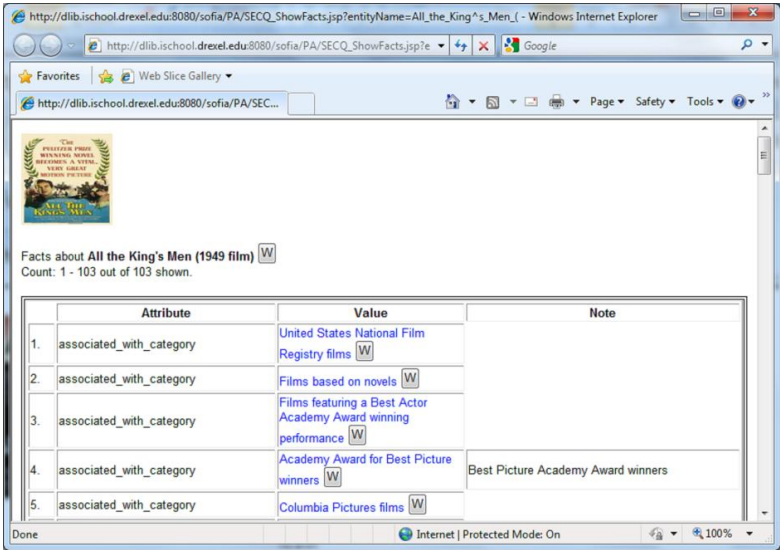
Results:


Count: 1 - 82 out of 82 shown.

	Image	Entity	Release	Attribute	Value	Note
1.		A Beautiful Mind (film) 	(2001)	has_won_award	Academy Award for Best Picture 	74th Academy Awards (2002-03-24) [for 2001] -- Producer: Brian Grazer; Ron Howard
2.		A Man for All Seasons (1966 film) 	(1966)	has_won_award	Academy Award for Best Picture 	39th Academy Awards (1967-04-10) [for 1966] -- Producer: Fred Zinnemann
3.		All About Eve 	(1950)	has_won_award	Academy Award for Best Picture 	23rd Academy Awards (1951-03-29) [for 1950] -- Producer: Darryl F. Zanuck
4.		All Quiet on the Western Front (1930 film) 	(1930)	has_won_award	Academy Award for Best Picture 	3rd Academy Awards (1930-11-05) [for 1929-1930] -- Producer: Carl Laemmle, Jr.
5.		All the King's Men (1949 film) 	(1949)	has_won_award	Academy Award for Best Picture 	22nd Academy Awards (1950-03-23) [for 1949] -- Producer: Robert Rossen
6.		Amadeus (film) 	(1984)	has_won_award	Academy Award for Best Picture 	57th Academy Awards (1985-03-25) [for 1984] -- Producer: Saul Zaentz
7.		American Beauty (film) 	(1999)	has_won_award	Academy Award for Best Picture 	72nd Academy Awards (2000-03-26) [for 1999] -- Producer: Bruce Cohen; Dan Jinks
8.		An American in Paris (film) 	(1951)	has_won_award	Academy Award for Best Picture 	24th Academy Awards (1952-03-20) [for 1951] -- Producer: Arthur Freed
9.		Annie Hall 	(1977)	has_won_award	Academy Award for Best Picture 	50th Academy Awards (1978-04-03) [for 1977] -- Producer: Charles H. Joffe
10.		Around the World in 80 Days (1956 film) 	(1956)	has_won_award	Academy Award for Best Picture 	29th Academy Awards (1957-03-21) [for 1956] -- Producer: Mike Todd
11.		Ben-Hur (1959 film) 	(1959)	has_won_award	Academy Award for Best Picture 	32nd Academy Awards (1960-04-04) [for 1959] -- Producer: Sam Zimbalist
12.		Braveheart 	(1995)	has_won_award	Academy Award for Best Picture 	68th Academy Awards (1996-03-25) [for 1995] -- Producer: Mel Gibson; Alan Ladd, Jr.; Bruce Davey
13.		Casablanca (film) 	(1942)	has_won_award	Academy Award for Best Picture 	16th Academy Awards (1944-03-02) [for 1943] -- Producer: Hal B. Wallis
14.		Cavalcade (1933 film) 	(1933)	has_won_award	Academy Award for Best Picture 	6th Academy Awards (1934-03-16) [for 1932-1933] -- Producer: Winfield Sheehan
15.		Chariots of Fire 	(1981)	has_won_award	Academy Award for Best Picture 	54th Academy Awards (1982-03-29) [for 1981] -- Producer: David Puttnam

Figure 82 PanAnthropon: GERQ function query result

If the user clicks on any entity name (highlighted in blue color) anywhere in the query result, a separate window showing all the facts on the entity (retrieved via SECQ function) appears, as shown in Figure 83. If the user clicks on  button that appears after the name of an entity in the Entity or Value field, a separate window for the Wikipedia page on the selected entity appears, as shown in Figure 84.



Facts about **All the King's Men (1949 film)** 
Count: 1 - 103 out of 103 shown.






	Attribute	Value	Note
1.	associated_with_category	United States National Film Registry films 	
2.	associated_with_category	Films based on novels 	
3.	associated_with_category	Films featuring a Best Actor Academy Award winning performance 	
4.	associated_with_category	Academy Award for Best Picture winners 	Best Picture Academy Award winners
5.	associated_with_category	Columbia Pictures films 	

Figure 83 PanAnthropon: GERQ function entity fact window



All the King's Men (1949 film) - Wikipedia, the free encyclopedia

Article Discussion Read Edit View history Search

All the King's Men (1949 film)
From Wikipedia, the free encyclopedia

This article is about the 1949 film. For other uses, see All the King's Men (disambiguation).

All the King's Men is a 1949 drama film based on the Robert Penn Warren novel of the same name. It was directed by Robert Rossen and starred Broderick Crawford in the role of Willie Stark.

Contents (hide)

- Plot
- Cast
- Production
- Awards
 - Academy Awards – 1949
- See also

All the King's Men

THE PULITZER PRIZE WINNING NOVEL BECOMES A VITAL, VERY GREAT MOTION PICTURE

Figure 84 PanAnthropon: GERQ function entity Wikipedia page window

Figure 85 shows the initial screen of the SECQ interface page.

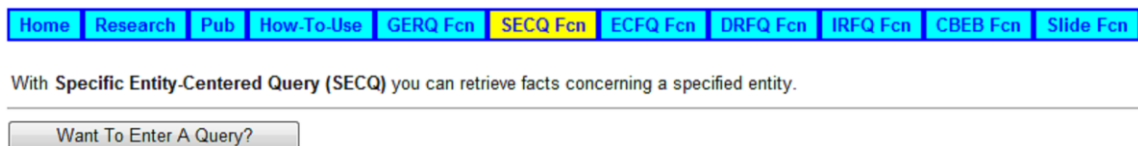
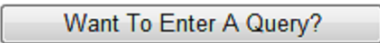


Figure 85 PanAnthropon: SECQ function interface initial screen

Once the user clicks on  button, an initial query form, shown in Figure 86, appears. Once the user selects an entity type and an entity subtype, similarly as in GERQ, either an input box for typing (partial) word(s) to get suggestions for an entity name appears, as shown in Figure 87, or a menu for entity name selection immediately appears, as shown in Figure 88, depending on the number of entities that match the specified entity type and entity subtype.

Select entity type, subtype, and name:

Select an entity type ▾

Figure 86 PanAnthropon: SECQ function initial query form

work ▾ film ▾

Name:

Figure 87 PanAnthropon: SECQ function query form after entity subtype selection #1

event ▾ film award event ▾

Name:

Figure 88 PanAnthropon: SECQ function query form after entity subtype selection #2

Once the user selects an entity name, a **SUBMIT QUERY** appears, as shown in Figure 89.

work film

Name: **The Lord of the Rings: The Return of the King**

SUBMIT QUERY

Figure 89 PanAnthropon: SECQ function query form after entity name selection

Once the user submits the query and once the query processing is completed, the result of the query appears, as partially shown in Figure 90. The query result is sorted in the alphabetical order of attribute names. Since SECQ is concerned with a single entity specified by the user, the result is presented in the form of <attribute, value, note>. As in GERQ, the user can click on any entity name or **W** button in the query result to open a window containing the facts or the Wikipedia page on the entity.

Query: Entity Type: work; Entity Subtype: film; Entity Name: The Lord of the Rings: The Return of the King

Results:



Facts about The Lord of the Rings: The Return of the King **W**
Count: 1 - 158 out of 158 shown.

	Attribute	Value	Note
1.	alt_name_url	The Lord of the Rings: The Return of the King (film) W	
2.	associated_with_category	Epic films W	
3.	associated_with_category	Sequel films W	
4.	associated_with_category	Academy Award for Best Director winning films W	Films whose director won the Best Director Academy Award
5.	associated_with_category	Academy Award for Best Adapted Screenplay winning films W	Films whose writer won the Best Adapted Screenplay Academy Award

Figure 90 PanAnthropon: SECQ function query result

Figure 91 shows the initial screen of the ECFQ interface page.



Figure 91 PanAnthropon: ECFQ function interface initial screen

Once the user clicks on button, an initial query form appears, as shown in Figure 92.

Select entity type, subtype, and names:

Select an entity type ▾

Figure 92 PanAnthropon: ECFQ function initial query form

Once the user selects an entity type and an entity subtype, similarly as in GERQ and SECQ, either an input box for typing words to get suggestions for entity 1 name appears, as shown in Figure 93, or a menu for entity 1 name selection immediately appears, similarly as in SECQ. Once the user selects a name for entity 1, then a similar process is repeated for the selection of the name for entity 2.

person ▾ any (director, actor, producer, etc.) ▾

Name of Entity 1:

Figure 93 PanAnthropon: ECFQ function query form after entity subtype selection

Once the user selects the names of both entity 1 and entity 2, a appears, as shown in Figure 94.

Name of Entity 1:

Name of Entity 2:

Figure 94 PanAnthropon: ECFQ function query form after entity 2 name selection

Once the user submits the query and once the query processing is completed, the result of the query appears, as partially shown in Figure 95. The query result is sorted in the alphabetical order of attribute names. Since ECFQ is concerned with commonalities between two specified entities, the result is presented in the form of <attribute, value, note_e1, note_e2>, where note_e1 and note_e2 represent (combined) notes for entity 1 and entity 2, respectively, for a common attribute–value pair.

Query: Entity Type: person; Entity Subtype: any; Entity 1 Name: Sean Connery; Entity 2 Name: Michael Caine

Results:

Commonalities between [Sean Connery](#)  and [Michael Caine](#) 
 Count: 1 - 11 out of 11 shown.




	Common Attribute	Common Value	Note for Entity 1	Note for Entity 2
1.	cast_member_of_film	The Man Who Would Be King (film) 	(1975)	(1975)
2.	cast_member_of_film	A Bridge Too Far (film) 	(1977) -- as Maj. Gen. Roy Urquhart	(1977) -- as Lt. Col. John Ormsby Evelyn "JOE" Vandeleur
3.	has_family_name_initial	C		
4.	has_won_award	Academy Award for Best Supporting Actor 	60th Academy Awards (1988-04-11) [for 1987] -- Film and Role: {{The Untouchables (film)}} as [[Jim Malone]]	59th Academy Awards (1987-03-30) [for 1986] -- Film and Role: {{Hannah and Her Sisters}} as [[Elliot]] 72nd Academy Awards (2000-03-26) [for 1999] -- Film and Role: {{The Cider House Rules (film)}} as [[Dr. Wilbur Larch]]
5.	has_won_award_in_event_month	01	45th Golden Globe Awards (1988-01-23) [for 1987] -- Golden Globe Award for Best Supporting Actor--Motion Picture -- Film and	41st Golden Globe Awards (1984-01-28) [for 1983] -- Golden Globe Award for Best Actor--Motion Picture Musical or Comedy -- Film and Role: {{Educating Rita (film)}} as [[Frank Bryant]]

Figure 95 PanAnthropon: ECFQ function query result

Figure 96 shows the initial screen of the DRFQ interface page.

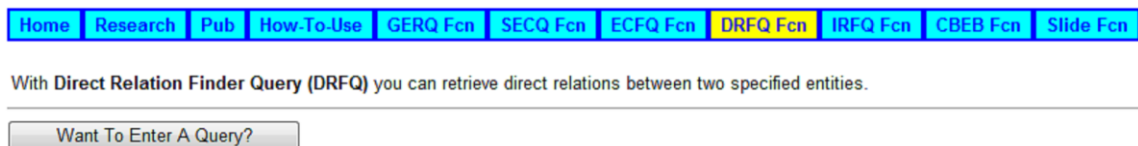
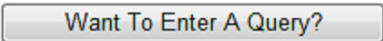


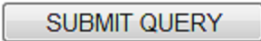
Figure 96 PanAnthropon: DRFQ function interface initial screen

Once the user clicks on  button, an initial query form appears, as shown in Figure 97.

Select entity type, subtype, and name for two entities:

Entity 1 Type/Subtype:

Figure 97 PanAnthropon: DRFQ function initial query form

Since DRFQ does not require that two entities specified in a query be of the same entity type and subtype, a menu for entity type selection again appears for entity 2, once the user selects the type, subtype, and name of entity 1, as shown in Figure 98. Once the user selects the type, subtype, and name for both entity 1 and entity 2, a  appears, as shown in Figure 99.

Entity 1 Type/Subtype:

Entity 1 Name:

Entity 2 Type/Subtype:

Figure 98 PanAnthropon: DRFQ function query form after entity 1 name selection

Entity 1 Type/Subtype:

Entity 1 Name:

Entity 2 Type/Subtype:

Entity 2 Name:

Figure 99 PanAnthropon: DRFQ function query form after entity 2 name selection

Once the user submits the query and once the query processing is completed, the result of the query appears, as partially shown in Figure 100. Since DRFQ is concerned with direct relations that hold between entity 1 and entity 2, the result is presented in the form of <entity_1, relation, entity_2, note>, where relation corresponds to an applicable attribute that connects entity 1 to entity 2.

Query: Entity 1 Type: person; Entity 1 Subtype: any; Entity 1 Name: Peter Jackson; Entity 2 Type: work; Entity 2 Subtype: film; Entity 2 Name: The Lord of the Rings: The Two Towers

Results:



Direct Relations between [Peter Jackson](#) and [The Lord of the Rings: The Two Towers](#)
 Count: 1 - 9 out of 9 shown.

	Entity 1	Relation	Entity 2	Note
1.	Peter Jackson	cameoed_in_film	The Lord of the Rings: The Two Towers	(2002) -- as Rohirrim Warrior [cameo] (uncredited)
2.	Peter Jackson	cast_member_of_film	The Lord of the Rings: The Two Towers	(2002) -- as Rohirrim Warrior [cameo] (uncredited)
3.	Peter Jackson	directed_film	The Lord of the Rings: The Two Towers	(2002)
4.	Peter Jackson	nominee_for_award_for_film	The Lord of the Rings: The Two Towers	60th Golden Globe Awards (2003-01-19) [for 2002] -- Golden Globe Award for Best Director
5.	Peter Jackson	nominee_for_award_for_film	The Lord of the Rings: The Two Towers	60th Golden Globe Awards (2003-01-19) [for 2002] -- Golden Globe Award for Best Motion Picture--Drama -- (as Director/Producer)
6.	Peter		The Lord of the Rings:	75th Academy Awards (2003-03-23) [for 2002] -- Academy Award for Best

Figure 100 PanAnthropon: DRFQ function query result

Figure 101 shows the initial screen of the IRFQ interface page.

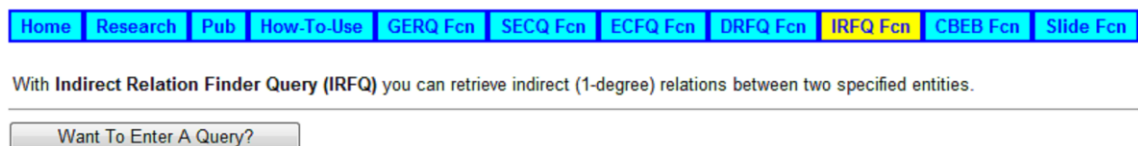


Figure 101 PanAnthropon: IRFQ function interface initial screen

The initial query form for IRFQ is identical to the one for DRFQ, as shown in Figure 102.

Select entity type, subtype, and name for two entities:

Entity 1 Type/Subtype:

Figure 102 PanAnthropon: IRFQ function initial query form

The process of query input for IRFQ is identical to that for DRFQ. Figures 103–105 show sample query results using IRFQ. As shown, the IRFQ function can be used to retrieve 1-degree indirect relations between a seemingly arbitrary pair of entities, such as Klaus Kinski and 16th Century.

Query: Entity 1 Type: **person**; Entity 1 Subtype: **any**; Entity 1 Name: **Klaus Kinski**; Entity 2 Type: **time**; Entity 2 Subtype: **century**; Entity 2 Name: **16th Century**

Results:

Indirect Relations between [Klaus Kinski](#)  and [16th Century](#)
Count: 1 - 2 out of 2 shown.

	Entity 1	E1-E3 Relation	Entity 3	E3-E2 Relation	Entity 2
1.	Klaus Kinski	cast_member_of_film	Aguirre, the Wrath of God 	set_in_(temporal_setting)	16th Century
2.	Klaus Kinski	starred_in_film	Aguirre, the Wrath of God 	set_in_(temporal_setting)	16th Century

Figure 103 PanAnthropon: IRFQ function query result #1

Query: Entity 1 Type: **concept**; Entity 1 Subtype: **film_subgenre**; Entity 1 Name: **SciFi/Fantasy-Related Film (theme/subgenre)**; Entity 2 Type: **person**; Entity 2 Subtype: **any**; Entity 2 Name: **Arnold Schwarzenegger**

Results:

Indirect Relations between **SciFi/Fantasy-Related Film (theme/subgenre)** and **Arnold Schwarzenegger**
 Count: 1 - 11 out of 11 shown.

	Entity 1	E1-E3 Relation	Entity 3	E3-E2 Relation	Entity 2
1.	SciFi/Fantasy-Related Film (theme/subgenre)	theme/subgenre_of_film	The Terminator <input type="button" value="W"/>	starring	Arnold Schwarzenegger
2.	SciFi/Fantasy-Related Film (theme/subgenre)	theme/subgenre_of_film	The Terminator <input type="button" value="W"/>	has_cast_member	Arnold Schwarzenegger
3.	SciFi/Fantasy-Related Film (theme/subgenre)	theme/subgenre_of_film	Predator (film) <input type="button" value="W"/>	starring	Arnold Schwarzenegger
4.	SciFi/Fantasy-Related Film (theme/subgenre)	theme/subgenre_of_film	Predator (film) <input type="button" value="W"/>	has_cast_member	Arnold Schwarzenegger
5.	SciFi/Fantasy-Related Film (theme/subgenre)	theme/subgenre_of_film	Total Recall <input type="button" value="W"/>	starring	Arnold Schwarzenegger
6.	SciFi/Fantasy-Related Film (theme/subgenre)	theme/subgenre_of_film	Total Recall <input type="button" value="W"/>	has_cast_member	Arnold Schwarzenegger
7.	SciFi/Fantasy-Related Film (theme/subgenre)	theme/subgenre_of_film	Terminator 2: Judgment Day <input type="button" value="W"/>	starring	Arnold Schwarzenegger
8.	SciFi/Fantasy-Related Film (theme/subgenre)	theme/subgenre_of_film	Terminator 2: Judgment Day <input type="button" value="W"/>	has_cast_member	Arnold Schwarzenegger

Figure 104 PanAnthropon: IRFQ function query result #2

Query: Entity 1 Type: **place**; Entity 1 Subtype: **city**; Entity 1 Name: **Philadelphia, Pennsylvania**; Entity 2 Type: **convention**; Entity 2 Subtype: **film_award**; Entity 2 Name: **Academy Award for Best Actor**

Results:

Indirect Relations between **Philadelphia, Pennsylvania** and **Academy Award for Best Actor**
 Count: 1 - 3 out of 3 shown.

	Entity 1	E1-E3 Relation	Entity 3	E3-E2 Relation	Entity 2
1.	Philadelphia, Pennsylvania	geographical_setting_for_film	Rocky <input type="button" value="W"/>	produced_nominee_for_award	Academy Award for Best Actor
2.	Philadelphia, Pennsylvania	geographical_setting_for_film	Witness (1985 film) <input type="button" value="W"/>	produced_nominee_for_award	Academy Award for Best Actor
3.	Philadelphia, Pennsylvania	geographical_setting_for_film	Philadelphia (film) <input type="button" value="W"/>	produced_winner_of_award	Academy Award for Best Actor


Figure 105 PanAnthropon: IRFQ function query result #3

As shown above, the result of the IRFQ query is presented without the `note` field either for the relation between entity 1 and entity 3 or for the relation between entity 3 and entity 2, in order to make it easier to see how the two specified entities, entity 1 and entity 2, are indirectly related via entity 3.

Figure 106 shows the initial screen of the CBEB interface page.



Figure 106 PanAnthropon: CBEB function interface initial screen

If the user clicks on the  button, an initial query form appears, as shown in Figure 107. The menu for top super-category selection contains three options, as shown in Figure 108.

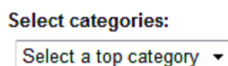


Figure 107 PanAnthropon: CBEB function initial query form

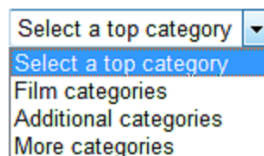


Figure 108 PanAnthropon: CBEB function menu for top super-category selection

Once the user selects a top-level super-category, a menu for subcategory selection appears, as shown in Figure 109. The menu contains those super-categories that have been classified under the selected top-level super-category, as shown in Figure 110. Once the user selects a sub-super-category from the menu, additional menus may appear for selection of sub-super-categories, until the user reaches the leaf level that represents the level of actual Wikipedia categories assigned to the entities (film pages), as shown in Figure 111. The menu for leaf-level category selection presents category names together with the number of associated films, as shown in Figure 112.

Film categories ▼

Select a subcategory ▼

Figure 109 PanAnthropon: CBEF function query form after top super-category selection

Select a subcategory

- Film categories by award
- Film categories by country
- Film categories by crew/company
- Film categories by fictional character/milieu
- Film categories by genre
- Film categories by language
- Film categories by plot source
- Film categories by setting
- Film categories by shooting location
- Film categories by specific film (/series/source/character)
- Film categories by specific series/franchise/remake/spinoff/parody
- Film categories by style
- Film categories by subject
- Film categories by technology
- Film categories by theme/subgenre
- Film categories by timeline
- Film categories by topic/character portrayed
- Film categories by type
- TV film/program categories
- Other film categories

Select a subcategory ▼

Figure 110 PanAnthropon: CBEF function menu for sub-super-category selection

Film categories ▼

Film categories by genre ▼

Action film categories ▼

Select a leaf-level category ▼

Figure 111 PanAnthropon: CBEF function query form at leaf category selection stage

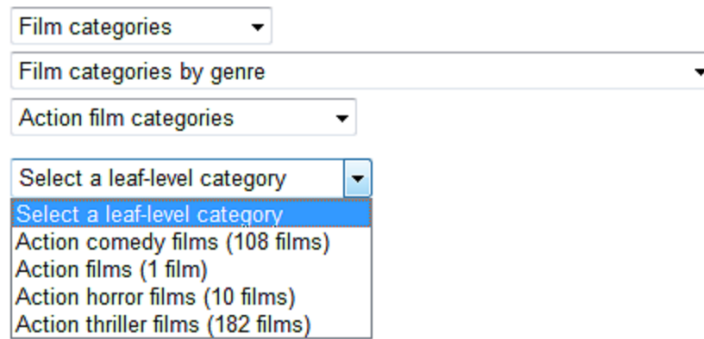


Figure 112 PanAnthropon: CBEB function menu for leaf category selection

Once the user selects a leaf category, query processing starts, as shown in Fig. 113. A partial snapshot of the query result is shown in Fig. 114.

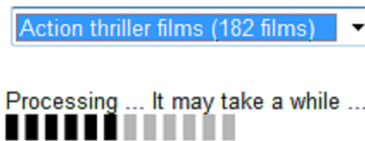


Figure 113 PanAnthropon: CBEB function query processing

Results:

Count: 1 - 182 out of 182 shown.

	Image	Title	Year	Wiki
1.		A Man Apart	2003	W
2.		Aces: Iron Eagle III	1992	W
3.		Air Force One (film)	1997	W
4.		Armored (film)	2009	W
5.		Assault on Precinct 13 (1976 film)	1976	W

Figure 114 PanAnthropon: CBEB function query result

Figure 115 shows the initial screen of the Slide function interface page.



Figure 115 PanAnthropon: Slide function interface initial screen

Once the user clicks on button, the interface starts retrieving the list of the titles of 11,355 films in the knowledge base, as shown in Figure 116.

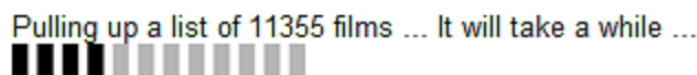


Figure 116 PanAnthropon: Slide function film list retrieval

Once the film title list is retrieved, a menu for film title selection appears, as shown in Figure 117. The menu contains the title of all films in the alphabetical order, as shown in Figure 118. Once the user selects a film title, brief introductory information concerning the film (extracted from the abstract section of the film page) is presented, together with the image of the film poster (if available), as shown in Figure 119. Again, the user can click on the film title or button in order to open the (SECQ) fact page on the film or the Wikipedia page on the film. The user can easily move forward/backward between film titles in the menu shown in Figure 118, to the effect of playing a seamless slide show consisting of information slides for each film.

Once you select a film name, you can use up/down arrow keys on your keyboard to move backward/forward between film titles.



Figure 117 PanAnthropon: Slide function query form after film list retrieval



Figure 118 PanAnthropon: Slide function menu for film name selection

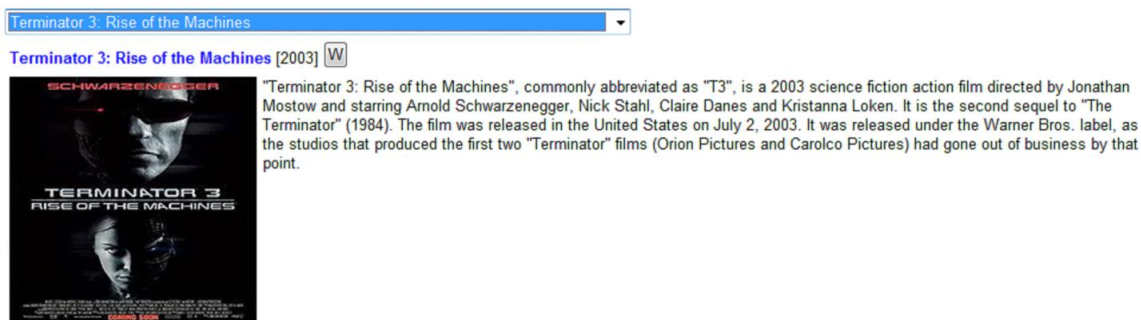


Figure 119 PanAnthropon: Slide function film info slide

7.3 Interface Usage Example

By using the interface constructed from this research, in particular, by using the main GERQ function, the user can issue sophisticated semantics-based queries, by explicitly specifying the entity type and subtype and by combining different sets of attributes and values, which is not possible to do on other Web sites. Figure 120 shows the GERQ query form representing a sample query with multiple conditions. Figure 121 shows the query result, which contains one entity, i.e., A Man for All Seasons (1966 film).

Select an entity type, subtype, and (up to 5) conditions:

work film

AND directed_by Fred Zinnemann

AND has_cast_member Paul Scofield

AND belongs_to_theme/subgenre_of Historical Film (theme/subgenre)

AND has_won_award Academy Award for Best Picture

AND produced_winner_of_award Academy Award for Best Actor

SUBMIT QUERY OR Add Another Condition OR Remove Last Condition

Figure 120 PanAnthropon: GERQ function query form for sample query

Results:

Count: 1 - 1 out of 1 shown.

Image	Entity	Release	Attribute	Value	Note
1. 	A Man for All Seasons (1966 film) 	(1966)	directed_by	Fred Zinnemann 	
			has_cast_member	Paul Scofield 	as Sir Thomas More
			belongs_to_theme/subgenre_of	Historical Film (theme/subgenre)	
			has_won_award	Academy Award for Best Picture 	39th Academy Awards (1967-04-10) [for 1966] -- Producer: Fred Zinnemann
			produced_winner_of_award	Academy Award for Best Actor 	Paul Scofield -- 39th Academy Awards (1967-04-10) [for 1966] -- Role: [[Sir Thomas More]]

Want To Enter Another Query?

Figure 121 PanAnthropon: GERQ function sample query result

CHAPTER 8: EVALUATION

8.1 Evaluation Methodology Overview

Two types of evaluation have been performed in order to evaluate (1) the quality of data extracted by the information extraction system of this research and (2) the effectiveness of information retrieval using the search interface constructed from this research.

The purpose of the first type of evaluation is to show that the data extracted/derived through this research is of high quality, compared against the source data in Wikipedia.

The quality of data, or the effectiveness of information extraction, is evaluated in terms of two criteria:

- (1) Precision: What percentage of the data that has been extracted is accurate?
- (2) Recall: What percentage of the data contained in the source has been extracted?

The two criteria are measured by using the equations shown in Figure 122, which are analogous to the standard equations to compute precision and recall (with respect to the results of page/document retrieval) in conventional information retrieval.

- **Eq. 1:** Precision = $\frac{\text{\# of data elements correctly extracted (for a given film)}}{\text{\# of data elements extracted (for a given film)}}$
- **Eq. 2:** Average Precision = $\frac{\sum_{i=1}^n \text{Precision}(i)}{n}$, $n = \text{\# of films in the test set}$
- **Eq. 3:** Recall = $\frac{\text{\# of data elements correctly extracted (for a given film)}}{\text{\# of data elements that should have been extracted (for a given film)}}$
- **Eq. 4:** Average Recall = $\frac{\sum_{i=1}^n \text{Recall}(i)}{n}$, $n = \text{\# of films in the test set}$

Figure 122 PanAnthropon: Equations for evaluation on information extraction

The purpose of the second type of evaluation is to show that the mechanism of retrieving answers by entity type/condition-specified queries (via the GERQ function of the interface) enables the user to issue more sophisticated queries and find the answers more directly and effectively than otherwise possible. It may be noted that the main intent of this type of evaluation is not to demonstrate the usability or user-friendliness of the search interface, as interpreted as ease or simplicity of use, but to demonstrate the effectiveness of information retrieval using the interface.

The effectiveness of information retrieval is evaluated in terms of precision and recall, computed by using the equations shown in Figure 123, which are analogous to the equations used for the evaluation on information extraction.

- **Eq. 5:** Precision = $\frac{\text{weighted \# of correct entities retrieved (for a given query)}}{\text{\# of all entities retrieved (for a given query)}}$
- **Eq. 6:** Average Precision = $\frac{\sum_{i=1}^n \text{Precision}(i)}{n}$, $n = \# \text{ of queries in the test set}$
- **Eq. 7:** Recall = $\frac{\text{weighted \# of correct entities retrieved (for a given query)}}{\text{\# of all correct entities (for a given query)}}$
- **Eq. 8:** Average Recall = $\frac{\sum_{i=1}^n \text{Recall}(i)}{n}$, $n = \# \text{ of queries in the test set}$

Figure 123 PanAnthropon: Equations for evaluation on information retrieval

The first type of evaluation concerning information extraction has been performed by manually inspecting a subset of the dataset constructed through this research and computing precision and recall with respect to the data contained in the test dataset. The second type of evaluation concerning information retrieval has been performed by conducting an experiment with human subjects that represent potential users and computing precision and recall with respect to the data collected from the experiment. Details concerning the two types of evaluation are provided in Section 8.2 and Section 8.3, respectively.

8.2 Evaluation: Information Extraction

8.2.1 Dataset

Considering the fact that the main source of information extraction/derivation in this research consisted of 10,640 Wikipedia pages on films, the test dataset for evaluation was constructed by retrieving the data extracted from a subset of 100 film pages. In order to evaluate the data quality in a balanced manner, 50 films in the test set were selected semi-randomly (by randomly choosing 50 films out of all films that have more than a set threshold number of film-centric facts) and the other 50 films were selected out of relatively well-known films. Film-centric facts (i.e., `<entity, attribute, value, note>` tuples where `entity` corresponds to a film) concerning each film in the 100-film set were retrieved and saved as a separate HTML page for each film to be compared with the source Wikipedia page.

8.2.2 Methods

The evaluation was performed by manually inspecting and comparing the facts extracted/derived about the 100 films in the test set against the facts (explicitly or implicitly) contained in (and intended to be extracted/derived from) the abstract, infobox, categories, and film cast information sections of the corresponding 100 source Wikipedia pages (saved in their condition at the time of information extraction). (Film-award-related facts, which were extracted/derived from the Wikipedia pages concerning the Academy Awards and Golden Globe Awards, were excluded from the evaluation.)

The `associated_with_category` facts extracted from the categories section were compared with the category links in the section. If the source page has a link to the category page “Epic films”, and if there is a corresponding fact `<film, associated_with_category, Epic films>` (where `film` stands for the title of a given film), then the extracted fact was considered

correct. If no such fact was extracted, then it was considered as a missing fact that failed to be extracted. If, on the other hand, the source page does not have such a category link as mentioned above but such a fact as described above was nevertheless extracted, then the extracted fact was considered as incorrect, in the sense that the source does not contain such a fact.

Additional facts that were indirectly derived from the categories, based on the taxonomy of super-categories, retain the ID of the corresponding source category in the `note` field, and, as such, it is possible to determine if such facts were correctly derived. For example, if the fact `<film, associated_with_category, Epic films>` was correctly extracted, and if the additional fact `<film, belongs_to_genre_of, Epic Film (genre)>` was derived from the former, then the latter was considered as a correct fact that was implicitly present in the source.

The facts extracted from the infobox section were compared with the attribute–value pairs in the infobox in the source page to determine correctness. These facts include those concerning the directors, producers, writers, narrators, starring actors, cinematographers, editors, musicians, studios, distributors, release dates, running times, countries, and languages associated with a film. If the source infobox contains, for example, the “Directed by” information field and “Victor Fleming” as one of its values, and if a fact `<film, directed_by, Victor Fleming>` was extracted, then the fact was considered correct.

The facts extracted from the abstract section include `also_known_as` facts and the facts concerning the directors, producers, writers, and starring actors (in case such information was not provided in the infobox). Both types of facts were checked against the source to determine their correctness.

The facts extracted or derived from the film cast information section include `has_cast_member` facts and `has_role` facts. If the film cast information section contains information about “Clark Gable” being a cast member as “Rhett Butler”, then the fact `<film, has_cast_member, Clark Gable, as Rhett Butler>` was considered as correct. Accordingly, the derivative fact `<film, has_role, Rhett Butler (role), played by Clark Gable>` was also considered correct. In case a source page does not contain information on the roles played by cast members, the correctness of the facts extracted was judged by considering only the actor names.

Based on the results of inspection and comparison as described above, precision and recall scores (in percentage) were first computed for each film in the test set individually (using *Eq. 1* and *Eq. 3*), and average precision and recall scores were then computed for all films as a whole (using *Eq. 2* and *Eq. 4*). (For the sake of computation of precision/recall scores, those (correct) facts that have been indirectly derived from the directly-extracted facts were considered as implicitly present in the source pages, so that the total number of facts extracted or derived for a given film would not exceed the total number of facts in the source page. An equal correctness/incorrectness score unit of 1 was used for each fact to compute precision and recall.)

8.2.3 Results

Table 47 shows the summary results of the evaluation in terms of the number of facts in source vs. number of facts extracted vs. number of facts correctly extracted. The number of facts extracted (11,495) represents the sum of the number of facts extracted or derived for each film in the 100-film test set. The number of facts correctly extracted (11,491) represents the number of facts that have been judged correct. The number of facts in source (11,509) represents the number of facts explicitly or implicitly contained in the source Wikipedia pages (only with respect to the types of facts intended to be extracted/derived).

Table 47 PanAnthropon: Summary results of info extraction evaluation: number of facts

# of Facts in Source	# of Facts Extracted	# of Facts Correctly Extracted
11,509	11,495	11,491

Table 48 shows the number of films in the test set for each distinct precision/recall score pair. The table shows that per-film precision/recall scores for 88 out of 100 films in the set were 100% precision and 100% recall. It also shows that only 12 films had recall scores less than 100% and that only 3 films had precision scores less than 100%. Table 49 shows the average precision and recall scores for the test set as a whole. As shown, the results confirm high data quality with 99.96% average precision and 99.84% average recall.

Table 48 PanAnthropon: Summary results of info extraction evaluation: precision/recall

Precision (in %)	Recall (in %)	# of Films
100.00%	100.00%	88
100.00%	99.35%	1
100.00%	99.23%	1
100.00%	99.17%	1
100.00%	99.15%	1
100.00%	99.12%	1
100.00%	99.09%	1
100.00%	98.20%	1
100.00%	97.54%	1
100.00%	97.46%	1
99.12%	99.12%	1
98.59%	98.59%	1
97.89%	97.89%	1

Table 49 PanAnthropon: Summary results of info extraction evaluation: average precision/recall

Average Precision	Average Recall	Total # of Films
99.96%	99.84%	100

Table 50 (continued on p.165) shows detailed evaluation results, with the title, numbers of source vs. extracted vs. correct facts, and precision/recall scores for each film.

Table 50 PanAnthropon: Detailed results of info extraction evaluation

	Film Title	# Src. Facts	# Ext. Facts	# Corr. Facts	Precision	Recall
1	<i>Gone with the Wind (film)</i>	131	131	131	100.00%	100.00%
2	<i>Intolerance (film)</i>	110	109	109	100.00%	99.09%
3	<i>A Midsummer Night's Dream (1935 film)</i>	88	88	88	100.00%	100.00%
4	<i>All Quiet on the Western Front (1930 film)</i>	78	78	78	100.00%	100.00%
5	<i>MASH (film)</i>	108	108	108	100.00%	100.00%
6	<i>Network (film)</i>	79	79	79	100.00%	100.00%
7	<i>The Last Picture Show</i>	72	72	72	100.00%	100.00%
8	<i>The Apartment</i>	72	72	72	100.00%	100.00%
9	<i>The Bonfire of the Vanities (film)</i>	265	265	265	100.00%	100.00%
10	<i>The Godfather Part III</i>	110	110	110	100.00%	100.00%
11	<i>Terms of Endearment</i>	83	83	83	100.00%	100.00%
12	<i>Heaven Can Wait (1978 film)</i>	78	78	78	100.00%	100.00%
13	<i>Apollo 13 (film)</i>	124	124	124	100.00%	100.00%
14	<i>Amistad (film)</i>	114	114	114	100.00%	100.00%
15	<i>Everyone Says I Love You</i>	115	115	115	100.00%	100.00%
16	<i>Dangerous Minds</i>	118	118	118	100.00%	100.00%
17	<i>United 93 (film)</i>	225	225	225	100.00%	100.00%
18	<i>X-Men: The Last Stand</i>	143	143	143	100.00%	100.00%
19	<i>One Flew Over the Cuckoo's Nest (film)</i>	99	99	99	100.00%	100.00%
20	<i>The Devil Wears Prada (film)</i>	96	96	96	100.00%	100.00%
21	<i>The Bourne Ultimatum (film)</i>	113	112	112	100.00%	99.12%
22	<i>The Red Violin</i>	118	117	117	100.00%	99.15%
23	<i>An Ideal Husband (1999 film)</i>	100	100	100	100.00%	100.00%
24	<i>Renaissance (film)</i>	99	99	99	100.00%	100.00%
25	<i>Li'l Abner (1959 film)</i>	82	82	82	100.00%	100.00%
26	<i>It's a Wonderful Life</i>	114	114	114	100.00%	100.00%
27	<i>Henry V (1944 film)</i>	118	118	118	100.00%	100.00%
28	<i>Rebecca (1940 film)</i>	95	95	95	100.00%	100.00%
29	<i>A Tale of Two Cities (1935 film)</i>	95	95	93	97.89%	97.89%
30	<i>The Longest Day (film)</i>	212	212	212	100.00%	100.00%
31	<i>Around the World in 80 Days (1956 film)</i>	192	192	192	100.00%	100.00%
32	<i>Lawrence of Arabia (film)</i>	130	129	129	100.00%	99.23%
33	<i>Giant (film)</i>	86	86	86	100.00%	100.00%
34	<i>Reds (film)</i>	125	125	125	100.00%	100.00%
35	<i>Goodfellas</i>	125	125	125	100.00%	100.00%
36	<i>The Deer Hunter</i>	107	107	107	100.00%	100.00%
37	<i>A Bridge Too Far (film)</i>	151	151	151	100.00%	100.00%
38	<i>The Lord of the Rings: The Return of the King</i>	128	128	128	100.00%	100.00%
39	<i>Gladiator (2000 film)</i>	99	99	99	100.00%	100.00%
40	<i>Alexander (film)</i>	136	136	136	100.00%	100.00%
41	<i>Patton (film)</i>	140	140	140	100.00%	100.00%
42	<i>2001: A Space Odyssey (film)</i>	111	111	111	100.00%	100.00%
43	<i>Rocky II</i>	153	152	152	100.00%	99.35%
44	<i>Star Trek (film)</i>	122	119	119	100.00%	97.54%
45	<i>Troy (film)</i>	119	119	119	100.00%	100.00%
46	<i>300 (film)</i>	114	114	113	99.12%	99.12%
47	<i>For Whom the Bell Tolls (film)</i>	69	69	69	100.00%	100.00%
48	<i>Doctor Zhivago (film)</i>	108	108	108	100.00%	100.00%
49	<i>All the King's Men (1949 film)</i>	64	64	64	100.00%	100.00%
50	<i>From Here to Eternity</i>	108	108	108	100.00%	100.00%

Table 50 (continued)

51	<i>The Citadel (film)</i>	105	105	105	100.00%	100.00%
52	<i>The Best Years of Our Lives</i>	75	75	75	100.00%	100.00%
53	<i>The Green Years (film)</i>	91	91	91	100.00%	100.00%
54	<i>The Thin Man (film)</i>	76	76	76	100.00%	100.00%
55	<i>It's a Mad, Mad, Mad, Mad World</i>	252	252	252	100.00%	100.00%
56	<i>Casino Royale (1967 film)</i>	136	136	136	100.00%	100.00%
57	<i>Giant (film)</i>	86	86	86	100.00%	100.00%
58	<i>Last Tango in Paris</i>	97	97	97	100.00%	100.00%
59	<i>Dave (film)</i>	145	145	145	100.00%	100.00%
60	<i>Hairspray (1988 film)</i>	136	136	136	100.00%	100.00%
61	<i>Missing (film)</i>	87	87	87	100.00%	100.00%
62	<i>This Is Spinal Tap</i>	117	117	117	100.00%	100.00%
63	<i>Zoolander</i>	203	203	203	100.00%	100.00%
64	<i>Scary Movie 3</i>	147	147	147	100.00%	100.00%
65	<i>The English Patient (film)</i>	87	87	87	100.00%	100.00%
66	<i>Big Fish</i>	103	103	103	100.00%	100.00%
67	<i>Babel (film)</i>	113	113	113	100.00%	100.00%
68	<i>Letters from Iwo Jima</i>	103	103	103	100.00%	100.00%
69	<i>Apocalypto</i>	112	112	112	100.00%	100.00%
70	<i>I'm Not There</i>	111	109	109	100.00%	98.20%
71	<i>Gentleman's Agreement</i>	71	71	70	98.59%	98.59%
72	<i>The Transformers: The Movie</i>	147	147	147	100.00%	100.00%
73	<i>That's Life! (film)</i>	101	101	101	100.00%	100.00%
74	<i>The Polar Express (film)</i>	102	102	102	100.00%	100.00%
75	<i>The Happy Time</i>	78	78	78	100.00%	100.00%
76	<i>The Godfather</i>	97	97	97	100.00%	100.00%
77	<i>The Phantom of the Opera (1925 film)</i>	131	131	131	100.00%	100.00%
78	<i>The Wizard of Oz (1939 film)</i>	102	102	102	100.00%	100.00%
79	<i>A Beautiful Mind (film)</i>	90	90	90	100.00%	100.00%
80	<i>The Godfather Part II</i>	185	185	185	100.00%	100.00%
81	<i>Shakespeare in Love</i>	69	69	69	100.00%	100.00%
82	<i>West Side Story (film)</i>	97	97	97	100.00%	100.00%
83	<i>Casino Royale (2006 film)</i>	112	112	112	100.00%	100.00%
84	<i>Rain Man</i>	144	144	144	100.00%	100.00%
85	<i>Apocalypse Now</i>	125	125	125	100.00%	100.00%
86	<i>Platoon (film)</i>	125	125	125	100.00%	100.00%
87	<i>Million Dollar Baby</i>	88	88	88	100.00%	100.00%
88	<i>Titanic (1997 film)</i>	137	137	137	100.00%	100.00%
89	<i>The Lord of the Rings: The Two Towers</i>	120	119	119	100.00%	99.17%
90	<i>Black Hawk Down (film)</i>	136	136	136	100.00%	100.00%
91	<i>The Lord of the Rings: The Fellowship of the Ring</i>	99	99	99	100.00%	100.00%
92	<i>A Room with a View (film)</i>	110	110	110	100.00%	100.00%
93	<i>The Queen (film)</i>	92	92	92	100.00%	100.00%
94	<i>The Manchurian Candidate (2004 film)</i>	117	117	117	100.00%	100.00%
95	<i>Forrest Gump (film)</i>	103	103	103	100.00%	100.00%
96	<i>Rambo (film)</i>	97	97	97	100.00%	100.00%
97	<i>Up (2009 film)</i>	118	115	115	100.00%	97.46%
98	<i>Gandhi (film)</i>	113	113	113	100.00%	100.00%
99	<i>Pulp Fiction (film)</i>	77	77	77	100.00%	100.00%
100	<i>Indiana Jones and the Kingdom of the Crystal Skull</i>	103	103	103	100.00%	100.00%

8.3 Evaluation: Information Retrieval

The experimental study proposal (protocol #19542) for this research was approved by the IRB at Drexel University on 2011-04-26. The approval documentation is provided in Appendix.

8.3.1 Experimental Design

The evaluation on information retrieval was intended to demonstrate the relative effectiveness of information retrieval using the interface constructed from this research, in comparison with one of existing search interfaces. Even though Wikipedia served as the original source of the data extracted through this research, and, accordingly, the evaluation on information extraction used the Wikipedia pages for the comparison to validate the quality of the data extracted, it was decided not to use the Wikipedia interface in the evaluation on information retrieval. The decision was based on the consideration for fairness, given the fact that Wikipedia is a general- or multi-domain information source and that, as such, the search result returned when using the Wikipedia site will include items that are not relevant to the film domain. Accordingly, the Internet Movie Database (IMDb) Web site was chosen instead, considering the fact that its interface allows the user to search the content of the largest film-related database and that it is one of the most popular sites on the Web, frequently used by many users who must be familiar with its features.

The main task of the experiment to evaluate information retrieval effectiveness required the subjects to find answers to two subsets of 5 test questions each, by using the IMDb interface and by using (the GERQ function of) the PanAnthropon interface, respectively. The decision to use one group of subjects to perform the task using both interfaces, instead of using two distinct groups of subjects to use one or the other interface, was based on the consideration that such a design would prevent the potential interference due to the different levels of experience and proficiency between subjects and that it would thus ensure the validity and fairness of evaluation.

According to the chosen experimental design, the analysis of the main task results involved computing precision/recall for each test question per subject, computing average precision/recall for each subset of questions per subject, computing average precision/recall per subject group as a whole, and analyzing the results in terms of the comparison between IMDb and PanAnthropon.

The hypotheses that were tested through the experiment are as follows:

- Hypothesis 1: Per-subject average precision/recall will be generally higher for the PanAnthropon subset (i.e., the subset of questions that the given subject answered by using the GERQ function of the PanAnthropon interface) than for the IMDb subset.
- Hypothesis 2: Per-group average precision/recall will be higher for the PanAnthropon subset than for the IMDb subset.

8.3.2 Experimental Procedures

A total of 33 voluntary subjects were recruited to participate in the experiment. (Demographic data concerning the subjects are provided in Subsection 8.3.3.) Due to the scheduling conflicts among the subjects, the experiment was conducted via multiple sessions, with 2 to 7 subjects each, over the course of four days between 2011-05-13 and 2011-05-19, inclusive. The procedures used for each experimental session (except signing of the informed consent form and the compensation receipt) are described below.

8.3.2.1 Pre-experimental-task procedures

The subjects were first introduced to the general purpose and methods of the study. The subjects were then directed to the “How-To-Use” page of the PanAnthropon interface, partially shown in Fig. 124. The subjects were asked to read the general background information and the usage instructions for the GERQ function of the interface. Once the subjects finished reading, they were

given two sample queries (which are simpler than the actual main task questions) to see if they could find answers using the GERQ interface. (The subjects were given approximately 5 minutes in total to read the instructions and try sample queries.) The subjects were then directed to the IMDb homepage. Most subjects except two or three of them were already familiar with the IMDb interface. Those who were unfamiliar were asked to try various search functions available on the interface. The subjects were instructed to use only the GERQ function of the PanAnthropon interface when performing the main task. They were instructed to freely use any search functions available on the IMDb interface. (The subjects were asked to open only one page on the Web browser for each interface, but the instructions were not followed consistently.)

This page contains information to help the user understand how to use the interface.

BACKGROUND INFORMATION

1. Introduction

This interface is created mainly for the sake of demonstrating semantics-based information retrieval, which is contrasted with familiar keyword-based information retrieval.

The phrase "semantics-based information retrieval" here means retrieving entities or facts concerning entities directly (instead of indirectly retrieving related pages/documents) based on the semantic knowledge stored in the knowledge base.

2. Definitions

Entity refers a thing of any kind. For example, **Sean Connery** is an entity, so is the film **The Man Who Would Be King (film)**.

Type of an entity refers to a generic class into which the given entity is classified. For example, the type of the entity Sean Connery is **person**, whereas the type of the entity The Man Who Would Be King (film) is **work**.

Subtype of an entity refers to a (most specific) subclass into which the entity is classified. For example, the subtype of the entity Sean Connery is **actor/actress**, whereas the subtype of the entity The Man Who Would Be King (film) is **film**.

Attribute refers to a property (predicate) associated with an entity. In general, attributes applicable to an entity are dependent on the type/subtype of the entity. For example, a film is associated with a director who directed it. In that sense, the attribute **directed_by** is applicable to a film entity, and relates the film entity with a director entity. Conversely, the attribute **directed_film** is applicable to a director entity, and relates the director entity with a film entity.

Value refers to the value of an attribute (for a given entity). Many of the attributes in this project are named in such a way that suggests the types of values. For example, the attribute **has_won_award** suggests that the corresponding value should be an award entity. The attribute **has_won_award_at_event**, on the other hand, suggests that the value should be an event entity. Similarly, the attribute **has_won_award_for_film** suggests that the value should be a film entity.

Figure 124 PanAnthropon: How-To-Use page

Once the subjects were ready to start the main task, they were given semi-randomly-assigned task codes and subject IDs. They were then asked to fill out a pre-experimental-task questionnaire shown in Figure 125. The pre-experimental-task procedures were completed with the subjects filling out the questionnaire.

Task Code:

Participant ID:

Are you a student?

Yes No

Are you an undergraduate or graduate student? (If not a student, select N/A.)

Undergraduate Graduate N/A

What is your current major? (If not a student, write N/A.)

What is your age?

What level of experience do you have in online information search?

Expert Intermediate Novice

How often do you search for information online?

Several times per day

About once per day

A few times per week

About once per week

Less frequently

Figure 125 PanAnthropon: Pre-experimental-task questionnaire

8.3.2.2 Experimental-task procedures

All subjects were given the same task set consisting of 10 questions, divided into two subsets of 5 questions each, as shown in Figure 126. (The actual task sheets given to the subjects consisted of 5 pages, with 2 questions on each page, with sufficient white spaces left for writing down the answers.) One half of subjects ($N=12$) first answered Subset 1 using IMDb and then Subset 2 using PanAnthropon; the other half ($N=12$) first answered Subset 1 using PanAnthropon and then Subset 2 using IMDb. (Note: The total number, 24, represents the number of subjects whose main task data have been included in the analysis of the results, as will be explained in Subsection 8.3.3.) (The instructions, “Use IMDb Interface” or “Use PanAnthropon Interface”, were clearly given before each subset on the task sheets.) Three variations of question ordering were used for each subset of questions in a comparable manner, as shown in Table 51. (The questions were re-labeled on the actual task sheets, according to the order in which the questions were presented per each distinct task code.) The subjects were instructed to spend no more than 5 minutes per each question when performing the task, resulting in the total task time of approximately 50 minutes.

Table 51 PanAnthropon: Experimental task codes and corresponding question orderings

Code	N	Interface 1	Interface 1 Question Set	Interface 2	Interface 2 Question Set
X-1	4	IMDb	Q1 » Q2 » Q3 » Q4 » Q5	PanAnthropon	Q6 » Q7 » Q8 » Q9 » Q10
Y-1	4	PanAnthropon	Q1 » Q2 » Q3 » Q4 » Q5	IMDb	Q6 » Q7 » Q8 » Q9 » Q10
X-2	4	IMDb	Q3 » Q1 » Q5 » Q4 » Q2	PanAnthropon	Q8 » Q6 » Q10 » Q9 » Q7
Y-2	4	PanAnthropon	Q3 » Q1 » Q5 » Q4 » Q2	IMDb	Q8 » Q6 » Q10 » Q9 » Q7
X-3	4	IMDb	Q5 » Q4 » Q3 » Q2 » Q1	PanAnthropon	Q10 » Q9 » Q8 » Q7 » Q6
Y-3	4	PanAnthropon	Q5 » Q4 » Q3 » Q2 » Q1	IMDb	Q10 » Q9 » Q8 » Q7 » Q6

8.3.2.3 Post-experimental-task procedures

Once the subjects completed the main experimental task, they were given a post-experimental-task questionnaire consisting of 8 questions, as shown in Fig. 127. (The actual questionnaire given to the subjects consisted of 3 pages, allowing room for detailed responses from the subjects.) The experimental session was concluded with the subjects filling out the questionnaire.

Test Set for Subjects' Experimental Task

Subset 1

- Q1. **Who** played **ALL** of the following roles: **Clem, Fox, and Hickey**? (Give me the name of the actor/actress *and* the titles of the films in which he/she played the roles.)
- Q2. **Who** won **BOTH** the Golden Globe Award for Best Actress—Motion Picture Drama **AND** the Golden Globe Award for Best Actress—Motion Picture Musical or Comedy?
- Q3. Which **films** produced winner of Academy Award for Best Director **AND** nominee for Academy Award for Best Actor? (Once you find the films, give me the titles of only those films released in/after 2000, together with their release years.)
- Q4. At which (Academy or Golden Globe) **award events** was Tommy Lee Jones nominated for awards? (Give me the names/dates of the events *and* the names of the awards *and* the titles of the films for which he was nominated.)
- Q5. Give me the title of the **film**, which is directed by Richard Attenborough, **AND** which has Sean Connery as a cast member, **AND** which belongs to the genre of War Epic Film, **AND** which is set in Netherlands, **AND** which has a role named Dr. Jan Spaander. (Give me *also* the role played by Sean Connery in the film *and* the name of the actor/actress who played the role Dr. Jan Spaander.)

Subset 2

- Q6. **Who** played **ALL** of the following roles: **Jim Malone, Robin Hood, and The Raisuli**? (Give me the name of the actor/actress *and* the titles of the films in which he/she played the roles.)
- Q7. **Who** won **BOTH** the Academy Award for Best Actor **AND** the Academy Award for Best Supporting Actor?
- Q8. Which **films** produced winner of Academy Award for Best Actor **AND** nominee for Academy Award for Best Director? (Once you find the films, give me the names of only those films released in/after 2000, together with their release years.)
- Q9. At which (Academy or Golden Globe) **award events** did Peter Ustinov win awards? (Give me the names/dates of the events *and* the names of the awards *and* the titles of the films for which he won the awards.)
- Q10. Give me the title of the **film**, which is directed by Werner Herzog, **AND** which has Klaus Kinski as a cast member, **AND** which belongs to the genre of Adventure Drama Film, **AND** which is set in (the) 16th Century, **AND** which has a role named Don Fernando de Guzman. (Give me *also* the role played by Klaus Kinski in the film *and* the name of the actor/actress who played the role Don Fernando de Guzman.)

Figure 126 PanAnthropon: Experimental task question set

Post-Experimental-Task Questionnaire

Q1. Overall, did you find the Internet Movie Database interface to be effective for searching for and retrieving answers to the queries? Please state **Yes**, **Maybe**, or **No**. Then please explain **why**.

Q2. Overall, did you find the PanAnthropon interface to be effective for searching for and retrieving answers to the queries? Please state **Yes**, **Maybe**, or **No**. Then please explain **why**.

Q3. **Which** of the two interfaces did you find to be **more** effective for finding answers?

Q4. Could you describe the reason for your answer to the previous question? (**Why** did you find the Internet Movie Database interface **OR** the PanAnthropon interface to be more effective for finding answers?)

Q5. Were you able to understand how to formulate queries and how to retrieve answers using the PanAnthropon interface? Please state **Yes**, **Maybe**, or **No**.

Q6. If you answered **Maybe** or **No** to the previous question, what kinds of difficulties did you have in understanding or using the PanAnthropon interface?

Q7. Would you be interested in using interfaces similar to the PanAnthropon interface to find answers to queries concerning the fields that you are interested in?

Q8. Any other comments concerning your experience with performing the test task?

Figure 127 PanAnthropon: Post-experimental-task questionnaire

8.3.3 Experimental Results

8.3.3.1 Pre-task questionnaire responses

The pre-experimental-task questionnaire has not been administered to 2 subjects. Out of 31 subjects that provided responses to the questionnaire, 30 subjects identified themselves as students. Tables 52–56 present the summary of the demographic data collected from 31 subjects.

Table 52 PanAnthropon: Pre-experimental-task questionnaire responses: student type

Student Type	N
UG Student	27
GR Student	3
N/A	1

Table 53 PanAnthropon: Pre-experimental-task questionnaire responses: student major

Major	N
Biomedical Engineering	7
Biology	6
Information Systems	5
Business Administration	5
Information Technology	2
Information Science	2
Business and Engineering	1
Electrical Engineering	1
Materials Engineering	1
N/A	1

Table 54 PanAnthropon: Pre-experimental-task questionnaire responses: age

Age	Min	18
	Max	45
	Average	21

Table 55 PanAnthropon: Pre-experimental-task questionnaire responses: info search experience

(Self-Assessed) Online Info Search Experience Level	N
Expert	13
Intermediate	17
Novice	1

Table 56 PanAnthropon: Pre-experimental-task questionnaire responses: info search frequency

Online Info Search Frequency	N
Several times per day	30
About once per day	1
A few times per week	0
About once per week	0
Less frequently	0

8.3.3.2 Main task results

All the questions in the main experimental task set have definite correct answers. Most questions, except Q2 and Q7, involve sub-questions that ask for relevant facts concerning the entities retrieved in response to the questions. To be considered completely correct, the answers to such questions should include the additional information requested. On the other hand, if an (element of the) answer to such a question contains all or some of the additional information requested but does not contain the correct main entity name, then such an (element of the) answer is considered completely wrong. Figures 128–130 show the weighted correctness scoring scheme used for each question in the main task set.

- Q1. **Who** played **ALL** of the following roles: **Clem, Fox,** and **Hickey**? (Give me the name, the actor/actress and the titles of the films in which he/she played the roles.)

(Number of Entities: 1)

Christopher Walken [70.00]
 played Clem in *Joe Dirt*. [10.00]
 played Fox in *New Rose Hotel*. [10.00]
 played Hickey in *Last Man Standing*. [10.00]

- Q2. **Who** won **BOTH** the Golden Globe Award for Best Actress—Motion Picture Drama **AND** the Golden Globe Award for Best Actress—Motion Picture Musical or Comedy?

(Number of Entities: 8)

Anne Bancroft [100.00]
 Julia Roberts [100.00]
 Marsha Mason [100.00]
 Meryl Streep [100.00]
 Nicole Kidman [100.00]
 Shirley MacLane [100.00]
 Sissy Spacek [100.00]
 Susan Hayward [100.00]

- Q3. Which **films** produced winner of Academy Award for Best Director **AND** nominee for Academy Award for Best Actor? (Once you find the films, give me the titles of only those films released on/after 2000, together with their release years.)

(Number of Entities: 4)

A Beautiful Mind [90.00]
 release: 2001 [10.00]
Brokeback Mountain [90.00]
 release: 2005 [10.00]
Million Dollar Baby [90.00]
 release: 2004 [10.00]
The Hurt Locker [90.00]
 release: 2008 [10.00]

Figure 128 PanAnthropon: Experimental task answer set #1

- Q4. At which (Academy or Golden Globe) **award events** was Tommy Lee Jones nominated for awards? (Give me the names/dates of the events and the names of the awards and the titles of the films for which he was nominated.)

(Number of Entities: 3)

38th Golden Globe Awards (or 1981 Golden Globe Awards) [70.00]

date: 1981-01-31 [10.00]

award: Golden Globe Award for Best Actor–Motion Picture Musical or Comedy [10.00]

film: *Coal Miner's Daughter* [10.00]

64th Academy Awards (or 1992 Academy Awards) [70.00]

date: 1992-03-30 [10.00]

award: Academy Award for Best Supporting Actor [10.00]

film: *JFK* [10.00]

80th Academy Awards (or 2008 Academy Awards) [70.00]

date: 2008-02-24 [10.00]

award: Academy Award for Best Actor [10.00]

film: *In the Valley of Elah* [10.00]

- Q5. Give me the title of the **film**, which is directed by Richard Attenborough, **AND** which has Sean Connery as a cast member, **AND** which belongs to the genre of War Epic Film, **AND** which is set in Netherlands, **AND** which has a role named Dr. Jan Spaander. (Give me also the role played by Sean Connery in the film and the name of the actor/actress who played the role Dr. Jan Spaander.)

(Number of Entities: 1)

A Bridge Too Far [80.00]

Sean Connery played Maj. Gen. Roy Urquhart. [10.00]

Dr. Jan Spaander was played by Laurence Olivier. [10.00]

- Q6. **Who** played **ALL** of the following roles: **Jim Malone**, **Robin Hood**, and **The Raisuli**? (Give me the name of the actor/actress and the titles of the films in which he/she played the roles.)

(Number of Entities: 1)

Sean Connery [70.00]

played Jim Malone in *The Untouchables*. [10.00]

played Robin Hood in *Robin and Marian*. [10.00]

played The Raisuli in *The Wind and the Lion*. [10.00]

- Q7. **Who** won **BOTH** the Academy Award for Best Actor **AND** the Academy Award for Best Supporting Actor?

(Number of Entities: 6)

Denzel Washington [100.00]

Gene Hackman [100.00]

Jack Lemmon [100.00]

Jack Nicholson [100.00]

Kevin Spacey [100.00]

Robert De Niro [100.00]

Figure 129 PanAnthropon: Experimental task answer set #2

- Q8. Which **films** produced winner of Academy Award for Best Actor **AND** nominee for Academy Award for Best Director? (Once you find the films, give me the names of only those films released on/after 2000, together with their release years.)

(Number of Entities: 6)

Capote [90.00]
 release: 2005 [10.00]
Gladiator [90.00]
 release: 2000 [10.00]
Milk [90.00]
 release: 2008 [10.00]
Mystic River [90.00]
 release: 2003 [10.00]
Ray [90.00]
 release: 2004 [10.00]
There Will Be Blood [90.00]
 release: 2007 [10.00]

- Q9. At which (Academy or Golden Globe) **award events** did Peter Ustinov win awards? (Give me the names/dates of the events and the names of the awards and the titles of the films for which he won the awards.)

(Number of Entities: 3)

33rd Academy Awards (or 1961 Academy Awards) [70.00]
 date: 1961-04-17 [10.00]
 award: Academy Award for Best Supporting Actor [10.00]
 film: *Spartacus* [10.00]
 37th Academy Awards (or 1965 Academy Awards) [70.00]
 date: 1965-04-05 [10.00]
 award: Academy Award for Best Supporting Actor [10.00]
 film: *Topkapi* [10.00]
 9th Golden Globe Awards (or 1952 Golden Globe Awards) [70.00]
 date: 1952-02-21 [10.00]
 award: Golden Globe Award for Best Supporting Actor–Motion Picture [10.00]
 film: *Quo Vadis* [10.00]

- Q10. Give me the title of the **film**, which is directed by Werner Herzog, **AND** which has Klaus Kinski as a cast member, **AND** which belongs to the genre of Adventure Drama Film, **AND** which is set in (the) 16th Century, **AND** which has a role named Don Fernando de Guzman. (Give me also the role played by Klaus Kinski in the film and the name of the actor/actress who played the role Don Fernando de Guzman.)

(Number of Entities: 1)

Aguirre, the Wrath of God [80.00]
 Klaus Kinski played Lope de Aguirre. [10.00]
 Don Fernando de Guzman was played by Peter Berling. [10.00]

Figure 130 PanAnthropon: Experimental task answer set #3

Main task data from 9 subjects have been excluded from analysis, due to various problems encountered during the experimental sessions, in order to ensure a valid assessment of the main task responses. Tables 57–60 show the summary results of the analysis of 24 subjects' task responses. As shown, the subjects' information retrieval task performance on the PanAnthropon interface clearly surpassed their performance on the IMDb interface, confirming both Hypothesis 1 and Hypothesis 2, despite the fact that the subjects had an extremely limited exposure to the PanAnthropon interface prior to performing the task.

Table 57 PanAnthropon: Experimental task results: per-group avg/max/min precision/recall

		PanAnthropon		IMDb	
		Score	N	Score	N
Average	Precision	83.11%		40.78%	
	Recall	83.55%		40.26%	
Max	Precision	100.00%	3	80.00%	1
	Recall	100.00%	5	78.00%	2
Min	Precision	58.00%	2	0.00%	3
	Recall	50.00%	1	0.00%	3

Table 58 PanAnthropon: Experimental task results: per-subject average precision/recall #1

	PanAnthropon		IMDb	
	N	%	N	%
Per-Subject Average 100% Precision 100% Recall	3	12.50%	0	0.00%
Per-Subject Average 0% Precision 0% Recall	0	0.00%	3	12.50%

Table 59 PanAnthropon: Experimental task results: per-subject average precision/recall #2

	PanAnthropon		IMDb	
	N	%	N	%
Per-Subject Average Precision > 90%	10	41.67%	0	0.00%
Per-Subject Average Recall > 90%	9	37.50%	0	0.00%

Table 60 PanAnthropon: Experimental task results: per-subject average precision/recall #3

	PanAnthropon		IMDb	
	N	%	N	%
Higher Per-Subject Average Precision	24	100.00%	0	0.00%
Higher Per-Subject Average Recall	24	100.00%	0	0.00%

Table 61 shows detailed experimental task results for each of 24 subjects.

Table 61 PanAnthropon: Detailed experimental task results

	Task Code	Subject ID	PanAnthropon Avg Precision	IMDb Avg Precision	PanAnthropon Avg Recall	IMDb Avg Recall
1	Y-1	8	98.00%	58.00%	98.00%	58.00%
2	Y-2	9	80.00%	0.00%	80.00%	0.00%
3	X-3	10	74.00%	50.80%	74.00%	58.00%
4	X-1	11	95.00%	58.00%	95.00%	58.00%
5	Y-3	12	100.00%	54.00%	100.00%	54.00%
6	X-3	13	60.00%	0.00%	50.00%	0.00%
7	Y-3	14	80.00%	40.00%	80.00%	33.33%
8	X-2	15	92.00%	0.00%	100.00%	0.00%
9	Y-2	16	88.00%	40.00%	100.00%	40.00%
10	X-3	17	78.00%	30.00%	78.00%	30.00%
11	Y-3	18	65.00%	18.00%	87.50%	18.00%
12	X-1	19	58.00%	18.00%	55.00%	18.00%
13	X-2	21	66.67%	40.00%	66.67%	22.50%
14	Y-2	22	100.00%	80.00%	100.00%	66.67%
15	X-3	23	94.00%	66.80%	90.67%	78.00%
16	Y-3	24	94.00%	34.00%	94.00%	34.00%
17	X-1	25	92.00%	59.60%	83.00%	54.50%
18	Y-1	26	78.00%	20.00%	78.00%	20.00%
19	X-2	27	58.00%	54.00%	58.00%	54.00%
20	Y-2	28	100.00%	47.33%	100.00%	57.33%
21	X-2	29	98.00%	64.13%	98.00%	78.00%
22	Y-1	30	78.00%	32.00%	78.00%	32.00%
23	X-1	31	96.00%	58.00%	89.33%	52.00%
24	Y-1	32	72.00%	56.00%	72.00%	50.00%

8.3.3.3 Post-task questionnaire responses

All 33 subjects' responses to the post-experimental-task questionnaire have been collected and analyzed. Tables 62 and 63 show the summary results of the analysis. As shown, 32 out of 33 subjects indicated an unwavering "Yes" in response to the question (Q2) on the effectiveness of the PanAnthropon interface for information retrieval. All 33 subjects unanimously agreed on the relative effectiveness of the PanAnthropon interface in comparison to the IMDb interface (Q3). Furthermore, despite the fact that the subjects were introduced to the new, unfamiliar interface for the first time in the experiment and that they were given a very limited amount of time to practice using it, 31 out of 33 subjects answered "Yes" to the question (Q5) on the understandability and usability of the PanAnthropon interface, with only 2 subjects expressing minor reservations. Furthermore, 29 out of 33 subjects indicated that they would be interested in using interfaces similar to the PanAnthropon interface for information retrieval tasks concerning the fields of their respective interests. The subjects' post-task questionnaire responses thus confirm, not only the relative effectiveness of the PanAnthropon interface, but also the usability of the interface. Given that the subject population represents typical information seekers who frequently engage in online search activities, it can be safely reasoned that the results are applicable to a broader population.

Table 62 PanAnthropon: Post-experimental-task questionnaire responses #1

	Yes		Maybe		No	
	N	%	N	%	N	%
Effectiveness of IMDb for Info Retrieval (Q1)	1	3.03%	7	21.21%	25	75.75%
Effectiveness of PanAnthropon for Info Retrieval (Q2)	32	96.97%	1	3.03%	0	0.00%
Understandability/Usability of PanAnthropon (Q5)	31	93.94%	2	6.06%	0	0.00%
Interest in Using Interfaces Similar to PanAnthropon (Q7)	29	87.88%	2	6.06%	2	6.06%

Table 63 PanAnthropon: Post-experimental-task questionnaire responses #2

	PanAnthropon		IMDb	
	N	%	N	%
Relative Effectiveness in Comparison (Q3)	33	100.00%	0	0.00%

The subjects' detailed responses to the post-experimental-task questionnaire indicate that most subjects found the PanAnthropon interface to be highly usable and effective, in itself and in comparison to the IMDb interface. Table 64 summarizes the main reasons given by the subjects concerning the usability and effectiveness of the PanAnthropon interface in terms of the design and configuration of the interface. Table 65 summarizes the main reasons in terms of the search function, process, and result.

Table 64 PanAnthropon: Reasons for effectiveness of PanAnthropon interface: design

No need to load multiple pages
Simplicity
Directness
Straightforwardness
Intuitiveness
Clear layout
Clear organization
Absence of excess graphics

Table 65 PanAnthropon: Reasons for effectiveness of PanAnthropon interface: function

No need to guess right keywords
Step-by-step search/query process
Easy understandability of entity types/subtypes and attributes
Availability of many attributes for specifying conditions
Capability to search for specific entities exactly
Capability to specify multiple conditions
Capability to issue complex, advanced queries
No need to browse multiple pages to find answers
Ease of making comparisons
Absence of extraneous information in query results

Tables 66–68 present the transcription of the subjects’ responses to the post-task questionnaire Q1 (“Overall, did you find the Internet Movie Database interface to be effective for searching for and retrieving answers to the queries? Please state **Yes**, **Maybe**, or **No**. Then please explain **why**.”), grouped by each response type. (The subject IDs shown in parentheses indicate that the main task data concerning those subjects have not been included in the analysis of the main experimental task responses.) Similarly, Tables 69 and 70 present the subjects’ responses to the post-task questionnaire Q2 (“Overall, did you find the PanAnthropon interface to be effective for searching for and retrieving answers to the queries? Please state **Yes**, **Maybe**, or **No**. Then please explain **why**.”).

Table 66 PanAnthropon: Yes (N=1) responses on effectiveness of IMDb

Subject ID	Task Code	Response
9	X-1	Yes. [Because] there were a lot of options to choose from. It is very detailed.

Table 67 PanAnthropon: Maybe (N=7) responses on effectiveness of IMDb

Subject ID	Task Code	Response
(3)	X-2	Maybe. It is ok. Sometimes it was hard to look for the answers and I had to open multiple IMDb pages to compare names, awards, etc and get to the answer.
(6)	Y-3	Maybe. The IMDb interface was effective for answering some of the queries. It was <u>not</u> effective for searching & retrieving answers to queries regarding [a]wards won. I was able to tailor queries regarding roles, actor/director names, etc. but could not do so with respect to awards received.
8	Y-1	Maybe. The IMDb is a good source, but it is hard to search for awards on there because the [interface] do[es]n’t include that option. The questions not related to awards were relatively easy to find.
12	Y-3	Maybe. For looking at the credentials of certain movies or actors, it was very easy. However, with respect to comparing multiple things, like finding an actor who won 2 awards, it was almost impossible.
22	Y-2	Maybe. The interface did not help in searching for specific information such as actors who won a certain award or films that have award-winning actors or directors. Queries regarding awards were difficult to answer.
26	Y-1	Maybe. Some [answers] were impossible to find[,] given the criterion needed[;] other searches just need[ed] the major keywords.
29	X-2	Maybe. It depends on what [I’]m looking for, and what [I’]m given. [I]f I was looking fo[r] award winners[,] it became difficult[,] but [for] most other [queries,] it was pretty easy.

Table 68 PanAnthropon: No (N=25) responses on effectiveness of IMDb

Subject ID	Task Code	Response
(1)	X-1	No. Because the advance[d] search function doesn't allow me to search different charact[e]rs who are acted by the same person. And it's kind of confusing to use the search function (too many attributes).
(2)	Y-1	No. There is no specific guideline as to searching specific information on awards, actors/actresses, films etc. Information is no[t] well organized.
(4)	Y-2	No. It's not meant to [and] doesn't support searches for separate things/fields. It doesn't have a completely searchable "Awards" section.
(5)	X-3	No. It took a long time since the search cannot be easily done. There are not many options.
(7)	X-1	No. I think there is too much information [and] garbage on the interface that makes it difficult to navigate and search.
10	X-3	No. The question[s] asked were very specific and these type[s] of questions were difficult to answer through IMDB.
11	X-1	No. IMDB was not effective in entering multiple queries. I had to employ a guess & check method to find my answers in IMDB.
13	X-3	No. The IMDB interface did not seem to be effective. This is because the 'advanced search' option for the IMDB did not provide any useful search mechanism. Some of the search options provided such as 'Name of Actor based on [H]eight' were unnecessary.
14	Y-3	No. The queries were specific like the intersection of two fields like award winner-actor and award winner-director. There is no way to [enter such queries and retrieve answers] in IMDB. IMDB appears to be more for casual browsing by title or name.
15	X-2	No. I stated in my Questionnaire why it was not satisfying. It was not intuitive, didn't provide a full range of searches.
16	Y-2	No. I could not find the awards on the website.
17	X-3	No. I think it [was] too long and didn't get me all the information needed.
18	Y-3	No. It was difficult finding specific answers.
19	X-1	No. The options were too specific and a person searching would not know that stuff beforehand.
(20)	X-1	No. Not that effective for detailed searches or for [the] majority of searches.
21	X-2	No. Because it was tough to find advanced search results.
23	X-3	No. The search bar did not allow for <u>multiple</u> parameters to be searched. It was a very basic search tool.
24	Y-3	No. The database contains a good wealth of information, however, there are limited ways to search through it and to narrow down by conditions.
25	X-1	No. The search functions on IMDB are not good for broad searches or very specific searches either. It was difficult to pinpoint the information.
27	X-2	No. Because it was t[e]dious.
28	Y-2	No. It required precise keywords to bring up "probable" answers. Then you had to search through those results to find right results.
30	Y-1	No. IMDB is useful for simple searches, for example[,] [a]ctors found in one film.
31	X-1	No. Because unless searching by actor or film it was near impossible to find the information. It had to be searched for in a round-about way.
32	Y-1	No. The questions stated asked for specific information, while IMDB is designed to provide a general overview on a film, person, etc. Not specific details.
(33)	X-1	No. It wasn't sensitive to the keywords I put in. I hardly got the answers I was looking for.

Table 69 PanAnthropon: Yes (N=32) responses on effectiveness of PanAnthropon

Subject ID	Task Code	Response
(1)	X-1	Yes. The user interface is simple, it doesn't include confusing words, it's easy to understand and use.
(2)	Y-1	Yes. "Entity" was easily identifiable, and organization [of information] was clear. Also [a] comparison would easily be made and [a] reference was readily available on Wikipedia.
(3)	X-2	Yes. It was really easy & straightforward to just select the categories & get the results in a list.
(5)	X-3	Yes. [Searching for and retrieving answers] could be more easily done [than on IMDB]. There were many options and the information could be retrieved easily using the search criteria.
(6)	Y-3	Yes. The PanAnthropon was effective (with some trial and error) to answer the queries. However, some of the sub-classes & value names were not intuitive – I imagine other queries would be more difficult if they relied on less intuitive values/attributes.
(7)	X-1	Yes. The organization of the interface allows for smooth navigation[,] making it easy to locate information.
8	Y-1	Yes. The PanAnthropon interface [provided] many different [query conditions] that can be used and [it] was just a matter of picking the right ones.
9	X-2	Yes. [Because] this [interface] allowed [me] to search [for] specific things and it was easier to locate [those specific things]. The only problem was the internet speed.
10	X-3	Yes. The questions were easy to answer through each variable.
11	X-1	Yes. PanAnthropon allowed me to enter multiple queries/conditions to easily find results.
12	Y-3	Yes. For the questions that were presented, it was easy to answer the questions because all of the comparisons were easy to make and [the answers were easy to] find.
13	X-3	Yes. The PanAnthropon interface was effective for searching for and retrieving answers to queries. This is because[:] a) [N]ot all the search [query conditions] [had to be] provided at once. It was a step[-]by[-]step process. b) [I]n order to change/modify [a query,] there was no need to hit the 'Back' button. It could be done in one page.
14	Y-3	Yes. As it seemed more comprehensive and advanced for performing logical operations like [AND], and had options for different [entity] types and subtypes. It is more advanced for querying.
15	X-2	Yes. As I became more familiar with the functionality of how the dropdown boxes looked and <u>more importantly</u> [as I became familiar with] the type[s] of filter[s] [found in those dropdown boxes], [I found that the interface] was very intuitive and allowed me to think of other possibilities relating to searching or other criteria/filters. [The interface] wasn't exciting to look at, but that wasn't the purpose of this.
16	Y-2	Yes. The layout was easy. I found everything with ease.
17	X-3	Yes. It was very good.
18	Y-3	Yes. It was easy finding answers to specific questions.
19	X-1	Yes. It was exact and had all the options that I needed.
(20)	X-1	Yes. Much better than IMDB. [I] can input more conditions. [It is] more specific.
21	X-2	Yes. It gave me many options up front.
22	Y-2	Yes. However, the interface is a bit cumbersome and slow. It takes a long time to load information and there are a lot of options to sort through when making a query.
23	X-3	Yes. This search engine allowed for multiple parameters and conditions to be found to narrow down the [query result] list very effectively. It was much more specific to what I was looking for.

Table 69 (continued)

24	Y-3	Yes. I was able to find the answers to all questions asked of me. The naming and organization of the attributes, however, was not always obvious and caused me a small amount of difficulty.
25	X-1	Yes. It was much easier to look for specific information as well as narrow down the information.
26	Y-1	Yes. Once all information was input for criteria needed, the info was there.
27	X-2	Yes. Because it gave key words to find things faster.
28	Y-2	Yes. Once you figured out how to utilize dropdown menus it was easy.
29	X-2	Yes. All you have to do is find the matching query and just [put] in what info you have and it gives you what you want.
30	Y-1	Yes. This interface allowed for more complicated searches while not being overly difficult. Albeit some of the entity types were confusing like “concept” vs “work”.
31	X-1	Yes. Because it focused on what you wanted. It searched for anything you were looking for in a direct straightforward way.
32	Y-1	Yes. The questions that were ask[ed] benefited from the search query [function] in the PanAnthropon interface due to the fact [that] the specific search query [function] had [the search criteria applicable to] those questions built in.
(33)	X-1	Yes. But I couldn't get some of the answers because of the time period.

Table 70 PanAnthropon: Maybe (N=1) response on effectiveness of PanAnthropon

Subject ID	Task Code	Response
(4)	Y-2	Maybe. The interface definitely needs some explanation and some keywords/conditions are confusing. But overall, it does produce answers based on the conditions you are looking for, as long as you know how to look for [them].

As mentioned, all 33 subjects indicated PanAnthropon in response to Q3 (“**Which** of the two interfaces did you find to be **more** effective for finding answers?”). Table 71 (continued on pp.185–186) presents the subjects’ responses to Q4 (“**Why** did you find the Internet Movie Database interface **OR** the PanAnthropon interface to be more effective for finding answers?”).

Table 71 PanAnthropon: Responses (N=33) on relative effectiveness of PanAnthropon vs. IMDB

Subject ID	Task Code	Response
(1)	X-1	The Internet Movie Database contains too much gra[ph]ic or unrelated information which confused me. PanAnthropon interface provides me with simple and understandable key words and attributes.
(2)	Y-1	Because information was well organized, each category was clearly separated[,] and it was very easy to make comparison[s]. All these features were not present in IMDB.

Table 71 (continued)

(3)	X-2	More straightforward, easy to find answers just by selecting categories down the list.
(4)	Y-2	PanAnthropon supports this type of “querying” while IMDB is more [for] a general search. If the questions were less specific, PanAnthropon might not be good since there seems no general search function.
(5)	X-3	The more advanced the search, the more useful it gets. The PanAnthropon search was useful in this regard.
(6)	Y-3	It was more effective at pinpointing items based on overlapping attributes. To me, it was comparable to conducting an “advanced search” in a database w/ a very specific thesaurus structure.
(7)	X-1	As stated earlier, the organization of the interface and the clarity of the interface [were] very helpful in locating/retrieving information.
8	Y-1	[On] PanAnthropon [it] was just [a matter of] picking the right query [conditions] and searching[,] which made it easier to find the answers.
9	X-2	IMDB did not have many options. It seemed disorganized.
10	X-3	For the type[s] of questions asked, PanAnthropon was able to more effectively answer the questions. It was much easier to isolate such answers.
11	X-1	The questions asked had multiple conditions. PanAnthropon allowed for inputting several conditions & finding only those outputs which satisfied <u>all</u> conditions. IMDB does not seem to have this characteristic.
12	Y-3	The questions gave us exactly what to click [from] the dropdown menus, making it much easier. The options pertained exactly to the questions. IMDB didn’t have any mechanism to compare movies.
13	X-3	The PanAnthropon interface provided a faster response to the queries. The IMDB on the other hand lacked search options for some important searches such as the Academy Awards.
14	Y-3	As the queries were more advanced and involved logical operations like AND and involved fields, types and subtypes, the PanAnthropon appeared to be more effective for them.
15	X-2	PanAnthropon site allowed more options [and] flexibility and allowed me to “Think” about other answers that I could want to know in the future.
16	Y-2	I found all answers quickly on that site.
17	X-3	Because I found most answers quicker [and] [it was] easier to use.
18	Y-3	Because it was a lot easier finding answers to specific questions.
19	X-1	[B]ecause it was easy to use and did not give you extraneous info.
(20)	X-1	More conditions. More inputs. Detail[ed] inquiries.
21	X-2	I found it easier because the interface allowed the user[s] to get what they need in an organized & effective manner.
22	Y-2	The questions being asked were exactly what that interface was made for. IMDB did not effectively answer those types of questions.
23	X-3	Much more detailed and very specific search methods [are provided by] PanAnthropon. IMDB is not specific to <u>multiple</u> parameters.
24	Y-3	PanAnthropon allowed you to filter by a much higher granularity and was able to present all of the query results in one place, rather than [making you have] to browse around from entity type to entity type, like you must do for IMDB.
25	X-1	I thought the PanAnthropon [interface] was easier to [mine data] through and find the specific information I wanted.
26	Y-1	Easier to input criteria and find the results [that are] more direct.
27	X-2	[B]ecause it made it easier to find things.
28	Y-2	It was quicker, simple & straightforward. It didn’t require you to guess the right keywords.

Table 71 (continued)

29	X-2	PanAnthropon is straight to the point. [Y]ou put in what you have and it gives you your answer. [B]ut with IMDB you have to search around.
30	Y-1	PanAnthropon allowed me to be more specific and to cross my searches[,] [u]nlike IMDB [where] I would have to open many tabs to answer questions [such as] which actors have had supporting roles and [which actors have won] [the] best actor award.
31	X-1	The PanAnthropon interface was more effective because it was searchable with any topic.
32	Y-1	The questions posed to us required a Boolean method of searching in which we needed to place logic[al] conditions in the search. PanAnthropon allowed for this while IMDB did not.
(33)	X-1	It was more specific and had more options.

As mentioned, 31 out of 33 subjects answered “Yes” to Q5 (“Were you able to understand how to formulate queries and how to retrieve answers using the PanAnthropon interface? Please state **Yes**, **Maybe**, or **No**.”). Table 72 shows the responses to Q6 (“If you answered **Maybe** or **No** to the previous question, what kinds of difficulties did you have in understanding or using the PanAnthropon interface?”) from 2 subjects who answered “Maybe” to Q5.

Table 72 PanAnthropon: Maybe (N=2) responses on usability of PanAnthropon interface

Subject ID	Task Code	Response
30	Y-1	The first few entity types are difficult to understand[,] but once you get the hang of what you [are] searching [for,] it is much easier. PanAnthropon makes [me] ask <u>what</u> or <u>who</u> I [am] looking for[.]
(33)	X-1	I had to use trial and error a lot to find the answers.

Tables 73–75 show responses from the subjects who answered “No” ($N=2$), “Maybe” ($N=2$), and “Yes” ($N=29$) to Q7 (“Would you be interested in using interfaces similar to the PanAnthropon interface to find answers to queries concerning the fields that you are interested in?”).

Table 73 PanAnthropon: No (N=2) responses on interest in using similar interfaces

Subject ID	Task Code	Response
10	X-3	Not really. [A]s a biology major I couldn’t really see the different values that make this [PanAnthropon interface] effective for films.
27	X-2	No.

Table 74 PanAnthropon: Maybe (N=2) responses on interest in using similar interfaces

Subject ID	Task Code	Response
(4)	Y-2	Perhaps.
(33)	X-1	Maybe.

Table 75 PanAnthropon: Yes (N=29) responses on interest in using similar interfaces

Subject ID	Task Code	Response
(1)	X-1	Yes. I like the user interface to be simple enough to use, [and does not] contain too much unrelated information.
(2)	Y-1	<u>YES.</u>
(3)	X-2	Yes. It makes it a lot easier [to find answers to queries]. :)
(5)	X-3	Yes. I would.
(6)	Y-3	Yes. It would be useful for identifying objects with specific, precise and unambiguous values that overlap.
(7)	X-1	Yes. It allows for information retrieval within [a] short period.
8	Y-1	Yes.
9	X-2	Yes. [Because] it makes it easy to locate specific things.
11	X-1	Yes. The interface is user-friendly and straightforward.
12	Y-3	Yes. It was much more convenient.
13	X-3	Yes. Having interfaces similar to the PanAnthropon interface would be useful. This is because it is easier to remember the context of where the name of the movie was heard [e.g.: Academy Award 2011] than the actual name itself. For such purposes the PanAnthropon [interface] was useful.
14	Y-3	Yes. It would help, especially in a professional environment. In a personal/home environment, [a] generic search is easier to use and works better.
15	X-2	Emphatically "Yes".
16	Y-2	Yes.
17	X-3	Yes. I really liked it.
18	Y-3	Yes. It is very easy finding answers to specific, highly detailed questions.
19	X-1	Yes.
(20)	X-1	<u>Yes.</u>
21	X-2	Yes. Outside [the domain] of movies as well.
22	Y-2	Yes. If I had questions like the ones in the study, the PanAnthropon interface is perfect for them.
23	X-3	Yes. Very useful and quick.
24	Y-3	Yes. Although I am not extremely interested in movies, this type of querying system would be amazing to have when searching for scholarly articles for research papers.
25	X-1	Yes. If the information I am digging through has many attributes.
26	Y-1	Yes. Very simple.
28	Y-2	Yes. If I am searching for something as detail[-]oriented as movies and actors.
29	X-2	Yes.
30	Y-1	Yes. I would! For example, in biology we read articles that are of a certain field (virology, for example), of a certain time (2000s), and occasionally from either two or more of the same authors. I could see an interface like this being very useful in the scenarios. I can also see this interface being useful in the arts. I would greatly enjoy [seeing] this interface used in other fields. PanAnthropon was both simple and effective.
31	X-1	Yes. Because I could find the exact information I was looking for.
32	Y-1	Yes.

CHAPTER 9: CONCLUSION

This dissertation research set out to enable a new mode of information retrieval that directly retrieves answers to user's queries, rather than documents that potentially contain the answers. The approach is concerned with retrieving entities, specified by semantic type, subtype, and conditions, based on semantic knowledge extracted from documents. The main contribution of this research consists in demonstrating the utility, feasibility, and effectiveness of entity retrieval by exploiting Wikipedia as a semantic knowledge source. The results of this research indicate that (1) ontology-based information extraction can be performed on both semi-structured and unstructured portions of documents, and highly accurate facts can be derived from the extracted information through recursive inferences; (2) based on the extracted semantic knowledge, a cascading interface can be constructed to guide users in a step-by-step manner towards the answers to their queries; and (3) users tend to prefer the entity-based interface for entity retrieval, and they can use the interface effectively to specify query conditions and retrieve answers.

In contrast to the other approaches that focused on extracting semantic information from Wikipedia, such as YAGO and DBpedia, this research did not only use structured templates and category structures but also used unstructured or non-standardized portions of Wikipedia pages which are far more difficult to properly process to extract information. The numerical comparison has also shown that this research has extracted more information relative to the size of the data source. Furthermore, the distinction between this research and other projects concerned with semantic information extraction using Wikipedia or similar data sources consists in the fact that indirect derivation of information, using the entity classes, super-categories, and the information directly extracted, played a major role in this research. The evaluation on information extraction has confirmed the quality of information extracted and derived through this research.

The interface constructed from this research radically departs from the traditional keyword-based search interface model. In the keyword search interface, the user can only enter a query in the form of a phrase or sentence consisting of one or more keywords. The search engine then decomposes the query phrase/sentence into individual keywords devoid of semantic context, and then it returns a list of pages/documents based on the matching of individual keywords with the individual words found in pages/documents. In contrast, the PanAnthropon interface allows the user to explicitly specify the type, subtype, and conditions to be satisfied by the things searched for, and in return it provides a list of the things that precisely match all the specified constraints, together with helpful contextual information. The PanAnthropon interface also stands out apart from existing semantic search interfaces such as the ones in YAGO and DBpedia. The latter interfaces allow the user to formulate a query in a semantics-based fashion, in the format of `<subject, predicate, object>` triple, but they do not provide any facilitating mechanism that can guide and help the user to effectively and efficiently formulate such a query in terms of entering/selecting an appropriate value for each field. As a result, the user has to discover, largely by trial and error, what can constitute a legitimate, meaningful query that can return valid results. In contrast, the PanAnthropon interface does not only allow the user to formulate a query in a semantics-based manner, but it facilitates the process in an intuitive, step-by-step fashion by providing only relevant menus and menu options according to the user selection in the previous step. Besides the fact the PanAnthropon interface provides several different semantic search functions, the major distinction between the PanAnthropon interface and the other semantic search interfaces consists in the querying/searching mechanism based on the explicit specification of entity type and subtype as well as `<attribute, value>` pairs. Contrary to a possible misconception that such a search mechanism may be beyond the capability of “ordinary” users, the evaluation experiment using potential users has not only shown that the subjects found the PanAnthropon interface to be highly effective in searching for specific things precisely, but it has

also shown that the subjects found the interface to be easily usable and understandable and that they were highly interested in using similar search interfaces.

To sum up, the main contribution from this research consists in demonstrating the advantage of the approach that explicitly utilizes a coherent classification of entities for the purpose of both effective information extraction and effective information retrieval. Additional contributions include the semantic dataset and interface themselves as resources that can be utilized beyond this research. The interface is already publicly accessible. The plan is under way to convert the dataset to different formats, e.g., XML, and to provide the resulting dataset as a freely-accessible resource for research purposes.

While the limitations of this research may consist in the fact that it used a single data source and a single domain of application, it is expected that the approaches used in this research can be readily applied to domains other than the film domain and other data sources on the Web that have semi-structured formats for presenting semantic information, with appropriate adjustments. The reason for such expectation is that, given the choice of domain and data source, the amount of effort required to apply the overall framework of this research — construction of a domain-oriented ontology, direct extraction of information, indirect derivation of information, and construction of a corresponding semantic search interface — is estimated to be within reasonable bounds. Of course, this is not to say that the exactly same extraction programs and the same interface can be used as such for other domains and data sources. However, as far as the application of the overall approaches is concerned, the cost of enabling such transference would be reasonable compared to the cost of restricting the user to the limited keyword-based search mechanism, as it is hoped that future works based on this research will be able to show.

Even though it was beyond of the scope of this dissertation research to further demonstrate that the approaches used in the research and the positive results thereby achieved can be generalized to other domains and data sources, I am hopeful that future work will demonstrate such a general applicability of the methods and outcomes of this research. To quote actual remarks from one of the subjects who participated in the experiment for this research, it is hoped and believed that “[this] study will make a significant impact [on] the future design and implementation of future search sites”.

List of References

- [Ada05] Adafre, S.F. and de Rijke, M. (2005). Discovering missing links in Wikipedia. In *Proceedings of the 2005 Workshop on Link Discovery: Issues, Approaches and Applications (LinkKDD 2005)* (Chicago, IL, USA, 21 August 2005). 2005.
- [Ahn04] Ahn, D., Jikoum, V., Mishne, G., Müller, K., de Rijke, M., and Schlobach, S. (2004). Using Wikipedia at the TREC QA Track. In *Proceedings of the 13th Text REtrieval Conference (TREC 2004)*. 2004.
- [Ath09a] Athenikos, S.J. (2009a). Interactive visualization and exploration of information on philosophers (and artists, scholars, & scientists) in an e-learning portal for digital humanities. Presented at the 2009 Symposium on Interactive Visual Information Collections and Activity (IVICA 2009) (Austin, TX, USA, 19 June 2009). Proceedings available at <http://www.csd.tamu.edu/~shipman/IVICA/2009proceedings.pdf>. Last accessed 30 July 2009.
- [Ath09b] Athenikos, S.J. (2009b). WikiPhiloSofia and PanAnthropon: extraction and visualization of facts, relations, and networks for a digital humanities knowledge portal. Poster. Presented at the ACM Student Research Competition at the 20th ACM Conference on Hypertext and Hypermedia (Hypertext 2009) (Torino, Italy, 29 June – 1 July 2009).
- [Ath10] Athenikos, S.J. (2010). Using Wikipedia to enable entity retrieval and visualization concerning the intellectual/cultural heritage . To appear in *Digital Humanities 2010 Conference Abstracts*. 2010.
- [AthL08a] Athenikos, S.J., and Lin, X. (2008a). The WikiPhil Portal: extraction, analysis, and visualization of philosophical connections using Wikipedia. Poster. Won student poster award at the Fall 2008 North East Database and Information Retrieval Day (University of Pennsylvania, Philadelphia, PA, USA, 14 October 2008).
- [AthL08b] Athenikos, S.J., and Lin, X. (2008b). The WikiPhil Portal: visualizing meaningful philosophical connections. Presented at the 2008 Chicago Colloquium on Digital Humanities and Computer Science (DHCS 2008) (University of Chicago, Chicago, IL, USA, 1–3 November 2008). In *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science*, 1, 1 (July 2009). 2009. Available at <https://letterpress.uchicago.edu/index.php/jdhcs/article/view/5/60>. Last accessed 30 July 2009.
- [AthL09a] Athenikos, S.J., and Lin, X. (2009a). WikiPhiloSofia: extraction and visualization of facts, relations, and networks concerning philosophers using Wikipedia. Presented at the 2009 Digital Humanities Conference (DH 2009) (University of Maryland, MD, USA, 22–25 June 2009). In *Digital Humanities 2009 Conference Abstracts*. 2009, pp. 56–62. Available at http://www.mith2.umd.edu/dh09/wp-content/uploads/dh09_conferencepreceedings_final.pdf. Last accessed 3 August 2009.
- [AthL09b] Athenikos, S.J., and Lin, X. (2009b). Visualizing intellectual connections among philosophers using the hyperlink & semantic data from Wikipedia. Poster. Presented at the 5th International Symposium on Wikis and Open Collaboration (WikiSym 2009) (Orlando, FL, USA, 25–27 October 2009).

- [Ats08] Atserias, J., Zaragoza, H., Ciaramita, M., and Attardi, G. (2008). Semantically annotated snapshot of the English Wikipedia. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)* (Marrakech, Morocco, 26 May – 1 June 2008). 2008, pp. 2313–2316.
- [Auer07a] Auer, S., and Lehmann, J. (2007). What have Innsbruck and Leipzig in common?: extracting semantics from wiki content. In *Proceedings of 4th European Semantic Web Conference (ESWC 2007)* (Innsbruck, Austria, 3–7 June 2007). 2007.
- [Auer07b] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. (2007). DBpedia: a nucleus for a Web of open data. In *Proceedings of the 6th International Semantic Web Conference (ISWC 2007) and the 2nd Asian Semantic Web Conference (ASWC 2007)* (Busan, South Korea, 11–15 November 2007), LNCS 4825. Springer-Verlag, Berlin/Heidelberg, 2007, pp. 722–735.
- [Baa03] Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (2003). *The Description Logic Handbook: Theory, Implementation and Applications*, Cambridge University Press, West Nyack, NY, USA, 2003.
- [Bai08] Bailey, P., Craswell, N., de Vries, A.P., and Soboroff. (2008). Overview of the TREC 2007 Enterprise Track. In *Proceedings of the Sixteenth Text REtrieval Conference (TREC 2007)*. Gaithersburg, MD, USA, 2008.
- [Bal09] Balasuriya, D., Ringland, N., Nothman, J., Murphy, T., and Curran, J.R. (2009). Named entity recognition in Wikipedia. In *Proceedings of the 2009 Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources, ACL-IJCNLP 2009* (Suntec, Singapore, 7 August 2009). ACL and AFNLP, 2009, pp. 10–18.
- [Ban07] Banko, M., Cafarella, M.J., Soderland, S., Broadhead, M., Etzioni, O. (2007). Open information extraction from the Web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI 2007)* (Hyderabad, India, 6–12 January 2007). 2007.
- [Bel05] Bellomi, F. and Bonato, R. (2005). Network analysis for Wikipedia. In *Proceedings of the 1st International Wikipedia Conference (Wikimania 2005)* (Frankfurt am Mein, Germany, 4–8 August 2005). 2005.
- [Ber01] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). The Semantic Web. *Scientific American*, 5 (May 2001).
- [Ber06] Berners-Lee, T. (2006). Linked data. 2006. Available at: <http://www.w3.org/DesignIssues/LinkedData.html>.
- [Bho07] Bhole, A., Fortuna, B., Grobelink, M., Mladenić, D. (2007). Mining Wikipedia and relating named entities over time. In *Proceedings of the Slovenian Conference on Data Mining and Data Warehouses (SiKDD 2007)* (Ljubljana, Slovenia, 12 October 2007). 2007.
- [Biz07] Bizer, C., Cyganiak, R., and Heath, T. (2007). How to publish linked data on the Web. 2007. Available at: <http://sites.wiwi.fu-berlin.de/suhl/bizer/pub/LinkedDataTutorial/>.

[Blo07] Blohm, S., and Cimiano, P. (2007). Using the Web to reduce data sparseness in pattern-based information extraction. In *Proceedings of the 11th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD 2007)* (Warsaw, Poland, 17–21 September 2007), LNCS 4702. Springer-Verlag: Berlin; Heidelberg, 2007, pp. 18–29.

[Bou07] Bouquet, P., Stoermer, H., and Giacomuzzi, D. (2007). OKKAM: enabling a Web of entities. In *Proceedings of the 16th International World Wide Web Conference (WWW 2007)* (Banff, Alberta, Canada, 8–12 May 2007). 2007.

[Bun06] Bunesu, R. and Paşca, M. (2006). Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)* (Trento, Italy, April 3–7, 2006). Association for Computer Linguistics, Morristown, NJ, USA, 2006.

[Bus06] Buscaldi, D., and Rosso, P. (2006). A Naïve bag-of-words approach to Wikipedia QA. In C. Peters et al. (eds.), *Evaluation of Multilingual and Multi-modal Information Retrieval, 7th Workshop of the Cross-Language Evaluation Forum (CLEF 2006)* (Alicante, Spain, 20–22 September 2006), Revised Selected Papers, LNCS 4730. Springer-Verlag, Berlin; Heidelberg, 2006, pp. 550–553.

[Bus07] Buscaldi, D., and Rosso, P. (2007). A comparison of methods for the automatic identification of locations in Wikipedia. In *Proceedings of the 4th Workshop on Geographic Information Retrieval (GIR 2007)* (Lisbon, Portugal, 9 November 2007). 2007.

[Cat08] Catone, J. (2008). Top 10 ways to search Wikipedia. 21 May 2008. Available at: http://www.readwriteweb.com/archives/top_10_ways_to_search_wikipedia.php. Last accessed 3 August 2009.

[Che07a] Cheng, T., Yan, X., and Chang, K.C.-C. (2007a). Supporting entity search: a large-scale prototype search system. In *Proceedings of ACM SIGMOD/PODS 2007 Conference (SIGMOD'07)* (Beijing, China, 11–14 June 2007). 2007.

[Che07b] Cheng, T., Yan, X., and Chang, K.C.-C. (2007b). EntityRank: search entities directly and holistically. In *Proceedings of the 33rd International Conference on Very Large Data Bases (VLDB'07)* (Vienna, Austria, 23–28 September 2007). VLDB Endowment, 2007, pp. 387–398.

[Cou09a] Coursey, K., Mihalcea, R. (2009a). Topic identification using Wikipedia graph centrality. In *Proceedings of NAACL HLT 2009: Short Papers* (Boulder, CO, USA, 31 May – 1 June 2009). Association for Computational Linguistics, Morristown, NJ, USA, 2009, pp. 117–120.

[Cou09b] Coursey, K., Mihalcea, R., and Moen, W. (2009b). Using encyclopedic knowledge for automatic topic identification. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL 2009)* (Boulder, CO, USA, 4–5 June 2009). 2009, pp. 210–218.

[Cra06] Craswell, N., de Vries, A.P., and Soboroff, I. (2006). Overview of the TREC-2005 Enterprise Track. In *Proceedings of the Fourteenth Text REtrieval Conference (TREC 2005)*. Gaithersburg, MD, USA, 2006.

- [Cra09] Craswell, N., Demartini, G., Gaugaz, J., and Iofciu, T. (2009). L3S at INEX 2008: retrieving entities using structured information. In Geva, S., Kamps, J., and Trotman, A. (eds.), *INEX 2008, LNCS 5631*. Springer-Verlag, Berlin; Heidelberg, 2009, pp. 253–263.
- [Cra01] Craswell, N., Hawking, D., Vercoustre, A.M., and Wilkins, P. (2001). P@noptic Expert: searching for experts not just for documents. In *Proceedings of the 7th Australian World Wide Web Conference (AusWeb01)* (Coffs Harbour, N.S.W., Australia, 2001). 2001, pp. 21–25.
- [Cuc07] Cucerzan, S. (2007). Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)* (Prague, Czech Republic, 28–30 June 2007). Association for Computational Linguistics, Morristown, NJ, USA, 2007, pp. 708–716.
- [Cui08] Cui, G., Lu, Q., Li, W., and Chen, Y. (2008). Corpus exploitation from Wikipedia for ontology construction. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)* (Marrakech, Morocco, 26 May – 1 June 2008). 2008, pp. 2125–2132.
- [Cul06] Culotta, A., McCallum, A., and Betz, J. (2006). Integrating probabilistic extraction models and data mining to discover relations and patterns in text. In *Proceedings of the 2006 Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL 2006)* (New York City, NY, USA, 4–9 June 2006). 2006.
- [deV08] de Vries, A.P., Vercoustre, A.-M., Thom, J.A., Craswell, N., and Lalmas, M. (2008). Overview of the INEX 2007 Entity Ranking Track. In Fuhr, N. et al. (eds.), *INEX 2007, LNCS 4862*. Springer-Verlag, Berlin/Heidelberg, 2008, pp. 245–251.
- [Dem09] Demartini, G., de Vries, A.P., Iofciu, T., and Zhu, J. (2009). Overview of the INEX 2008 Entity Ranking Track. In Geva, S., Kamps, J., and Trotman, A. (eds.), *INEX 2008, LNCS 5631*. Springer-Verlag, Berlin/Heidelberg, 2009, pp. 243–252.
- [Dem08] Demartini, G., Firan, C.S., and Iofciu, T. (2008). L3S at INEX 2007: query expansion for entity ranking using a highly accurate ontology. In Fuhr, N. et al. (eds.), *INEX 2007, LNCS 4862*. Springer-Verlag, Berlin/Heidelberg, 2008, pp. 252–263.
- [Den06] Denoyer, L., and Gallinari, P. (2006). The Wikipedia XML corpus. *SIGIR Forum*, 40(1) (June 2006), pp. 64–69.
- [Fel98] Fellbaum, C.E. (1998). *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, USA, 1998.
- [Fu07] Fu, L., Wang, H., Zhu, H., Zhang, H., Wang, Y., and Yong, Y. (2007). Making more Wikipedians: facilitating semantic reuse for Wikipedia authoring. In *Proceedings of the 6th International Semantic Web Conference (ISWC 2007) and the 2nd Asian Semantic Web Conference (ASWC 2007)* (Busan, South Korea, 11–15 November 2007), *LNCS 4825*. Springer-Verlag, Berlin/Heidelberg, 2007, pp. 128–141.

- [Gan09] Gantner, Z., and Schmidt-Thieme, L. (2009). Automatic content-based categorization of Wikipedia articles. In *Proceedings of the 2009 Workshop on the People's Web Meets NLP, ACL-IJCNLP 2009* (Suntec, Singapore, 7 August 2009). ACL and AFNLP, 2009, pp. 32–37.
- [Gre06] Gregorowicz, A. and Kramer, M.A. (2006). Mining a large-scale term-concept network from Wikipedia. Technical Report. MITRE Corporation, Bedford, MA, USA, 2006.
- [Gri09a] Grineva, M., Grinev, M., and Lizorkin, D. (2009a). Effective extraction of thematically grouped key terms from text. In *Proceedings of the AAAI 2009 Spring Symposium Series (AAAI-SS 2009)* (Stanford, CA, USA, 23–24 March 2009). 2009.
- [Gri09b] Grineva, M., Grinev, M., and Lizorkin, D. (2009b). Extracting key terms from noisy and multi-theme documents. In *Proceedings of the 18th International World Wide Web Conference (WWW 2009)* (Madrid, Spain, 20–24 April 2009). 2009, pp. 661–670.
- [Has09] Hassanzadeh, O. and Consens, M. (2009). Linked movie data base. In *Proceedings of the WWW 2009 Workshop on Linked Data on the Web (LDOW 2009)* (Madrid, Spain, 20 April 2009). 2009.
- [Hof09] Hoffmann, R., Amershi, S., Patel, K., Wu, F., Fogarty, J., and Weld, D.S. (2009). Amplifying community content creation with mixed-initiative information extraction. In *Proceedings of the 27th CHI Conference (CHI 2009)* (Boston, MA, USA, 4–9 April 2009). 2009.
- [Ift08] Iftene, A., and Balahur-Dobrescu, A. (2008). Named entity relation mining using Wikipedia. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)* (Marrakech, Morocco, 26 May – 1 June 2008). 2008, pp. 763–766.
- [Jäm08] Jämsen, J., Näppilä, T., and Arvola, P. (2008). Entity ranking based on category expansion. In Fuhr, N. et al. (eds.), *INEX 2007, LNCS 4862*. Springer-Verlag, Berlin/Heidelberg, 2008, pp. 264–278.
- [Jan07] Janik, M., and Kochut, K. (2007). Wikipedia in action: ontological knowledge in text categorization. Technical Report. University of Georgia Computer Science Department, Athens, GA, USA. Technical report no. UGA-CS-TR-07-001. 2007.
- [Jia09] Jiang, J., Lu, W., Rong, X., and Gao, Y. (2009). Adapting language modeling methods for expert search to rank Wikipedia entities. In Geva, S., Kamps, J., and Thom, J.A. (eds.), *INEX 2008, LNCS 5631*. Springer-Verlag, Berlin/Heidelberg, 2009, pp. 264–272.
- [Kai08] Kaiser, M. (2008). The QuALiM question answering demo: supplementing answers with paragraphs drawn from Wikipedia. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2008: HLT)* (Columbus, OH, USA, 15–20 June 2008). Association for Computational Linguistics, Morristown, NJ, USA, 2008.
- [Kap09] Kaptein, R., and Kamps, J. (2009). Finding entities in Wikipedia using links and categories. In Geva, S., Kamps, J., and Trotman, A. (eds.), *INEX 2008, LNCS 5631*. Springer-Verlag, Berlin/Heidelberg, 2009, pp. 273–279.

- [Kas08] Kasneci, G., Suchanek, F.M., Ifrim, G., Ramanath, M., and Weikum, G.** (2008). NAGA: searching and ranking knowledge. In *Proceedings of the 24th IEEE International Conference on Data Engineering (ICDE 2008)* (Cancun, Mexico, 7-12 April 2008). 2008.
- [Kaz07] Kazarma, J. and Torisawa, K.** (2007). Exploiting Wikipedia as external knowledge for named entity recognition. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)* (Prague, Czech Republic, 28–30 June 2007). Association for Computational Linguistics, Morristown, NJ, USA, 2007, pp. 698–707.
- [Kr07] Krötzsch, M., Schaffert, S., and Vrandečić, D.** (2007). Reasoning in semantic wikis. In *Proceedings of the Third International Summer School on Reasoning Web (Reasoning Web 2007)* (Dresden, Germany, 3–7 September 2007), *LNCS 4636*. Springer-Verlag, Berlin/Heidelberg, 2007, pp. 310–329.
- [Kr05] Krötzsch, M., Vrandečić, D., and Völkel, M.** (2005). Wikipedia and the Semantic Web - the missing links. In *Proceedings of the First International Wikimedia Conference (Wikimania 2005)* (Frankfurt, Germany, 4–8 August 2005). 2005.
- [Liz09] Lizorkin, D., Medelyan, O., and Grineva, M.** (2009). Analysis of community structure in Wikipedia. Poster Paper. In *Proceedings of the 18th International World Wide Web Conference (WWW 2009)* (Madrid, Spain, 20–24 April 2009). 2009, pp. 1221–1222.
- [Mat98] Mattox, D.** (1998). Expert Finder. *The Edge: The MITRE Advanced Technology Newsletter*, 2(1) (June 1998).
- [Med08a] Medelyan, O., and Legg, C.** (2008). Integrating Cyc and Wikipedia: folksonomy meets rigorously defined common-sense. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy (WikiAI 2008)* (Chicago, IL, USA, 13–14 July 2008). AAAI Press, 2008.
- [Med08b] Medelyan, O., and Milne, D.** (2008). Augmenting domain-specific thesauri with knowledge from Wikipedia. In *Proceedings of the New Zealand Computer Science Research Student Conference 2008 (NZCSRSC 2008)* (Christchurch, New Zealand, April 2008). 2008.
- [Med08c] Medelyan, O., Witten, I.H., and Milne, D.** (2008). Topic indexing with Wikipedia. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy (WikiAI 2008)* (Chicago, IL, USA, 13–14 July 2008). AAAI Press, 2008.
- [Mih07] Mihalcea, R.** (2007). Using Wikipedia for automatic word sense disambiguation. . In *Proceedings of the 2007 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2007)* (Rochester, NY, USA, 22–27 April 2007). Association for Computational Linguistics, Morristown, NJ, USA, 2007.
- [Mil07] Milne, D.** (2007). Computing semantic relatedness using Wikipedia link structure. In *Proceedings of the 5th New Zealand Computer Science Research Student Conference (NZCSRSC'07)* (Hamilton, New Zealand, 10–13 April 2007). 2007.

- [Mil06] Milne, D., Medelyan, O., and Witten, I.H. (2006). Mining domain-specific thesauri from Wikipedia: a case study. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006)* (Hong Kong, China, 18–22 December 2006). IEEE Computer Society, 2006, pp. 442–448.
- [Mil08a] Milne, D. and Witten, I.H. (2008a). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy (WikiAI 2008)* (Chicago, IL, USA, 13–14 July 2008). AAAI Press, 2008.
- [Mil08b] Milne, D. and Witten, I.H. (2008b). Learning to link with Wikipedia. In *Proceedings of the ACM 17th Conference on Information and Knowledge Management (CIKM 2008)* (Napa Valley, CA, USA, 26–30 October 2008). 2008.
- [Mil07a] Milne, D., Witten, I.H., and Nichols, D.M. (2007a). Extracting corpus specific knowledge bases from Wikipedia. Working Paper. Department of Computer Science, University of Waikato, New Zealand, June 2007.
- [Mil07b] Milne, D., Witten, I.H., and Nichols, D.M. (2007b). A knowledge-based search engine powered by Wikipedia. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM 2008)* (Lisbon, Portugal, 6–8 November 2007). ACM Press, New York, NY, 2007, pp. 445–454.
- [Mur08] Murugesan, M.S., and Mukherjee, S. (2008). An n-gram and initial description based approach for entity ranking track. In Fuhr, N. et al. (eds.), *INEX 2007, LNCS 4862*. Springer-Verlag, Berlin/Heidelberg, 2008, pp. 293–305.
- [Nak07a] Nakayama, K., Hara, T., and Nishio, S. (2007a). Thesaurus construction method from large scale Web dictionaries. In *Proceedings of the IEEE 21st International Conference on Advanced Information Networking and Applications (AINA 2007)* (Niagara Falls, Canada, 21–24 May 2007). 2007, pp. 932–939.
- [Nak07b] Nakayama, K., Hara, T., and Nishio, S. (2007b). Wikipedia mining for an association thesaurus construction. In *Proceedings of the 8th International Conference on Web Information Systems Engineering (WISE 2007)* (Nancy, France, 2–7 December 2007), LNCS 4831. Springer-Verlag, Berlin/Heidelberg, 2007.
- [Nak08b] Nakayama, K., Hara, T., and Nishio, S. (2008b). Wikipedia link structure and text mining for semantic relation extraction: towards a huge scale global Web ontology. In *Proceedings of the 5th European Semantic Web Conference (ESWC 2008)* (Tenerife, Spain, 1–5 June 2008). 2008.
- [Nak08c] Nakayama, K., Pei, M., Erdmann, M., Ito, M., Shirakawa, M., Hara, T., and Nishio, S. (2008). Wikipedia mining: Wikipedia as a corpus for knowledge extraction. In *Proceedings of the 4th Annual Wikipedia Conference (Wikimania 2008)* (Alexandria, Egypt, 17–19 July 2008). 2008.

- [Nas08a] **Nastase, V.** (2008). Topic-driven multi-document summarization with encyclopedic knowledge and spreading-activation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing (EMNLP 2008)* (Honolulu, HI, USA, 25–27 October 2008). 2008.
- [Ngu07] **Nguyen, D.P.T., Matsuo, Y., and Ishizuka, M.** (2007). Relation extraction from Wikipedia using subtree mining. In *Proceedings of the 22nd AAAI National Conference on Artificial Intelligence (AAAI 2007)* (Vancouver, British Columbia, Canada, 22–26 July 2007). AAAI Press, Menlo Park, CA, USA, 2007.
- [Not08] **Nothman, J., Curran, J.R., and Murphy, T.** (2008). Transforming Wikipedia into named entity training data. In *Proceedings of the ALTA 2008 Workshop*. 2008.
- [Ove09] **Overell, S., Sigurbjörnsson, B., and van Zwol, R.** (2009). Classifying tags using open content resources. In *Proceedings of the 2nd ACM International Conference on Web Search and Data Mining (WSDM 2009)* (Barcelona, Spain, 9–12 February 2009). 2009.
- [Ped08] **Pedro, V.C., Niculescu, R.S., and Lita, L.V.** (2008). Okinet: automatic extraction of a medical ontology from Wikipedia. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artificial Intelligence: An Evolving Synergy (WikiAI 2008)* (Chicago, IL, USA, 13–14 July 2008). AAAI Press, 2008.
- [Pic07] **Picca, D. and Popescu, A.** (2007). Using Wikipedia and supersense tagging for semi-automatic complex taxonomy construction. In *Proceedings of the 2007 RANLP Workshop on Computer-Aided Language Processing (CALP 2007)*. 2007.
- [Pon09] **Ponzetto, S.P. and Navigli, R.** (2009). Large-scale taxonomy mapping for restructuring and integrating Wikipedia. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence (IJCAI 2009)* (Pasadena, CA, USA, 11–17 July 2009). 2009.
- [Pon06] **Ponzetto, S.P., and Strube, M.** (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2006)* (New York, NY, USA, 4–9 June 2006). Association for Computational Linguistics, Morristown, NJ, USA, 2006, pp. 192–199.
- [Pon07a] **Ponzetto, S.P., and Strube, M.** (2007a). Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd National Conference on Artificial Intelligence (AAAI 2007)* (Vancouver, Canada, July 2007). AAAI Press, Menlo Park, CA, 2007, pp. 1440–1445.
- [Pon07b] **Ponzetto, S.P., and Strube, M.** (2007b). Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, 30 (2007), pp. 181–212.
- [Rah08] **Rahukar, M.A., Roth, D., and Huang, T.S.** (2008). Which “apple” are you taking about? Poster Paper. In *Proceedings of the 17th International World Wide Web Conference (WWW 2008)* (Beijing, China, 21–25 April 2008). 2008, pp. 1197–1198.

- [Rei08] **Reiter, N., Hartung, M., and Frank, A.** (2008). Resource-poor approach for linking ontology classes to Wikipedia articles. In *Proceedings of 2008 Symposium on Semantics in Systems for Text Processing (STEP 2008)* (Venice, Italy, 22–24 September 2008). 2008, pp.381–387.
- [Rui05] **Ruiz-Casado, M., Alfonseca, E., and Castells, P.** (2005). Automatic assignment of Wikipedia encyclopedic entries to Wordnet synsets. In *Proceedings of the 3rd Atlantic Web Intelligence Conference (AWIC 2005)* (Lodz, Poland, 6–9 June 2005). 2005.
- [Rui06] **Ruiz-Casado, M., Alfonseca, E., and Castells, P.** (2006). From Wikipedia to semantic relationships: a semi-automated annotation approach. In *Proceedings of the 3rd European Semantic Web Conference (ESWC 2006)* (Budva, Montenegro, 11–14 June 2006). 2006.
- [Rui07] **Ruiz-Casado, M., Alfonseca, E., and Castells, P.** (2007). Automising the learning of lexical patterns: an application to the enrichment of WordNet by extracting semantic relationships from Wikipedia. *Data & Knowledge Engineering*, 61(3) (June 2007), pp. 484–499.
- [Sar09] **Sarjant, S., Legg, C., Robinson, M., and Medelyan, O.** (2009). “All You Can Eat” ontology-building: feeding Wikipedia to Cyc. In *Proceedings of the 2009 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2009)* (Milan, Italy, 15–18 September 2009). 2009.
- [Sau09] **Sauper, C. and Barzilay, R.** (2009). Automatically generating Wikipedia articles: a structure-aware approach. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP (ACL-IJCNLP 2009)* (Suntec, Singapore, 2–7 August 2009). 2009, pp. 208–216.
- [Sch06] **Schönhofen, P.** (2006). Identifying document topics using the Wikipedia category network. In *Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006)* (Hong Kong, China, 18–22 December 2006). IEEE Computer Society, 2006, pp. 456–462.
- [Sch08] **Schönhofen, P.** (2008). Annotating documents by Wikipedia concepts. *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2008)* (Sydney, Australia, 9–12 December 2008). IEEE Computer Society, 2008, pp. 461–467.
- [Shi08] **Shiozaki, H., and Eguchi, K.** (2008). Entity ranking from annotated text collections using multitype topic models. In Fuhr, N. et al. (eds.), *INEX 2007, LNCS 4862*. Springer-Verlag, Berlin/Heidelberg, 2008, pp. 279–292.
- [Shn09] **Shnarch, E., Barak, L., and Dagan, I.** (2009). Extracting lexical reference rules from Wikipedia. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP (ACL-IJCNLP 2009)* (Suntec, Singapore, 2–7 August 2009). ACL and AFNLP, 2009, pp. 450–458.
- [Sob07] **Soboroff, I., de Vries, A.P., and Craswell, N.** (2007). Overview of the TREC 2006 Enterprise Track. In *Proceedings of the Fifteenth Text REtrieval Conference (TREC 2006)*. Gaithersburg, MD, USA, 2007.

[Str06] Strube, M., and Ponzetto, S.P. (2006). WikiRelate!: computing semantic relatedness using Wikipedia. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 2006)* (Boston, MA, USA, 10–16 July 2006). AAAI Press, Menlo Park, CA, 2006, pp. 1419–1424.

[Suc06a] Suchanek, F.M., Ifrim, G., and Weikum, G. (2006a). LEILA: learning to extract information by linguistic analysis. In *Proceedings of the 2nd Workshop on Ontology learning and Population: Bridging the Gap between Text and Knowledge (OLP2) at COLING/ACL 2006* (Sydney, Australia, 22 July 2006). 2006.

[Suc06b] Suchanek, F.M., Ifrim, G., and Weikum, G. (2006b). Combining linguistic and statistical analysis to extract relations from Web documents. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2006)* (Philadelphia, PA, USA, 20–23 August 2006). 2006.

[Suc07] Suchanek, F.M., Kasneci, G., and Weikum, G. (2007). YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. In *Proceedings of the 16th International World Wide Web Conference (WWW 2007)* (Banff, Alberta, Canada, 8–12 May 2007). ACM Press, New York, NY, 2007, pp. 697–706.

[Suc08] Suchanek, F.M., Kasneci, G., and Weikum, G. (2008). YAGO: a large ontology from Wikipedia and WordNet. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(3) (September 2008), pp. 203–217.

[Suc09] Suchanek, F.M., Sozio, M., and Weikum, G. (2009). SOFIE: a self-organizing framework for information extraction. In *Proceedings of the 18th International World Wide Web Conference (WWW 2009)* (Madrid, Spain, 20–24 April 2009). 2009, pp. 631–640.

[Sye07] Syed, Z.S., Finin, T., Joshi, A. (2007). Wikipedia as an ontology for describing documents. In *Proceedings of the 1st International Conference on Weblogs and Social Media (ICWSM 2007)* (Boulder, CO, USA, 26–28 March 2007). 2007.

[Tho08] Thomas, C., Mehra, P., Brooks, R., and Sheth, A. (2008). Growing fields of interest: using an expand and reduce strategy for domain model extraction. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2008)* (Sydney, Australia, 9–12 December 2008). 2008.

[Tor06] Toral, A. and Muñoz, R. (2006). A proposal to automatically build and maintain gazetteers for named entity recognition by using Wikipedia. In *Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics (EACL 2006)* (Trento, Italy, April 3–7, 2006). Association for Computer Linguistics, Morristown, NJ, USA, 2006.

[Tsi08] Tsikrika, T., Serdyukov, P., Rode, H., Westerveld, T., Aly, R., Hiemstra, D., and de Vries, A.P. (2008). Structured document retrieval, multimedia retrieval, and entity ranking using PF/Tijah. In Fuhr, N. et al. (eds.), *INEX 2007, LNCS 4862*. Springer-Verlag, Berlin/Heidelberg, 2008, pp. 306–320.

- [Ver08] **Vercoustre, A.-M., Pehcevski, J., and Thom, J.A.** (2008). Using Wikipedia categories and links in entity ranking. In Fuhr, N. et al. (eds.), *INEX 2007, LNCS 4862*. Springer-Verlag, Berlin/Heidelberg, 2008, pp. 321–335.
- [Völ06] **Völkel, M., Krötzsch, M., Vrandečić, D., Haller, H., and Studer, R.** (2006). Semantic Wikipedia. In *Proceedings of the 15th International World Wide Web Conference (WWW 2006)* (Edinburgh, Scotland, 23–26 May 2006). ACM Press, New York, NY, 2006, pp. 585–594.
- [Vra06] **Vrandečić, D., and Krötzsch, M.** (2006). Reusing ontological background knowledge in semantic wikis. In *Proceedings of the First Workshop on Semantic Wikis (SemWiki 2006) at the 3rd European Semantic Web Conference (ESWC 2006)* (Budva, Montenegro, 11–16 November 2006). 2006, pp. 16–30.
- [WaG07] **Wang, G., Yu, Y., and Zhu, H.** (2007). PORE: positive-only relation extraction from Wikipedia text. In *Proceedings of the 6th International Semantic Web Conference (ISWC 2007) and the 2nd Asian Semantic Web Conference (ASWC 2007)* (Busan, South Korea, 11–15 November 2007), *LNCS 4825*. Springer-Verlag, Berlin/Heidelberg, 2007, pp. 580–594.
- [WaP07] **Wang, P., Hu, J., Zeng, H.-J., Chen, L., and Chen, Z.** (2007). Improving text classification by using encyclopedia knowledge. In *Proceedings of the 7th IEEE International Conference on Data Mining (ICDM 2007)* (Omaha, NE, USA, 28–31 October 2007). IEEE Computer Society, 2007, pp. 332–341.
- [Wea06] **Weale, T.** (2006). Utilizing Wikipedia categories for document classification. Unpublished.
- [Wee09] **Weerkamp, W., Balog, K., and Meij, E.** (2009). A generative language modeling approach for ranking entities. In Geva, S., Kamps, J., and Trotman, A. (eds.), *INEX 2008, LNCS 5631*. Springer-Verlag, Berlin/Heidelberg, 2009, pp. 292–299.
- [Wei05] **Weischedel, R. and Brunstein, A.** (2005). *BBN Pronoun Coreference and Entity Type Corpus*. Linguistic Data Consortium, Philadelphia, 2005.
- [Wel08] **Weld, D.S., Wu, F., Adar, E., Amershi, S., Fogarty, J., Hoffmann, R., Patel, K., and Skinner, M.** (2008). Intelligence in Wikipedia. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI 2008)* (Chicago, IL, USA, 13–17 July 2008). 2008, pp. 1609–1614.
- [Wu07] **Wu, G., and Weld, D.S.** (2007). Autonomously semantifying Wikipedia. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM 2007)* (Lisbon, Portugal, 6–9 November 2007). ACM Press, New York, NY, 2007, pp. 41–50.
- [Wu08] **Wu, G., and Weld, D.S.** (2008). Automatically refining the Wikipedia infobox ontology. In *Proceedings of the 17th International World Wide Web Conference (WWW 2008)* (Beijing, China, 21–25 April 2008). 2008, pp. 635–644.
- [Yam08] **Yamangil, E., and Nelken, R.** (2008). Mining Wikipedia revision histories for improving sentence compression. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL 2008: HLT)* (Columbus, OH, USA, 15–20 June 2008). Association for Computational Linguistics, Morristown, NJ, USA, 2008.

[Yan09] Yan, Y., Okazaki, N., Matsuo, Y., Yang, Z., and Ishizuka, M. (2009). Unsupervised relation extraction by mining Wikipedia texts using information from the Web. In *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP (ACL-IJCNLP 2009)* (Suntec, Singapore, 2–7 August 2009). ACL and AFNLP, 2009, pp. 1021–1029.

[Yang07] Yang, J., Han, J., Oh, I., and Kwak, M. (2007). Using Wikipedia technology for topic maps design. In *Proceedings of the 45th ACM Annual Southeast Regional Conference (ACMSE 2007)* (Winston-Salem, NC, USA, 23–24 March 2007). Curan Associates, Inc., 2007, pp. 106–110.

[Yeh09] Yeh, E., Ramage, D., Manning, C.D., Agirre, E., and Soroa, A. (2009). WikiWalk: random walks on Wikipedia for semantic relatedness. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing, ACL-IJCNLP 2009* (Suntec, Singapore, 7 August, 2009). ACL and AFNLP, 2009, pp. 41–49.

[Yu07] Yu, J., Thom, J.A., and Tam, A. (2007). Ontology evaluation using Wikipedia categories for browsing. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM 2007)* (Lisbon, Portugal, 6–8 November 2007). 2007, pp. 223–232.

[Zar07] Zaragoza, H., Rode, H., Mika, P., Atserias, J., Ciaramita, M., and Attardi, G. (2007). Ranking very many typed entities on Wikipedia. In *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management (CIKM'07)* (Lisbon, Portugal, 6–8 November 2007). ACM Press, New York, NY, 2007, pp. 1015–1018.

[Zes07a] Zesch, T. and Gurevych, I. (2007). Analysis of the Wikipedia category graph for NLP applications. In *Proceedings of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies 2009 Conference (NAACL HLT 2009)* (Boulder, CO, USA, 31 May – 5 June 2009). Association for Computational Linguistics, Morristown, NJ, USA, 2009.

[Zes07b] Zesch, T., Gurevych, I., and Mühlhäuser, M. (2007). Analyzing and accessing Wikipedia as a lexical semantic resource. In *Proceedings of the Biannual Conference of the Society for Computational Linguistics and Language Technology* (Tübingen, Germany, 2007). 2007, pp. 213–221.

[Zha09] Zhang, Z. and Iria, J. (2009). Novel approach to automatic gazetteer generation using Wikipedia. In *Proceedings of the 2009 Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources, ACL-IJCNLP 2009* (Suntec, Singapore, 7 August 2009). 2009, pp. 1–9.

[Zhu08] Zhu, J., Song, D., and Rüger, S. (2008). Integrating document features for entity ranking. In Fuhr, N. et al. (eds.), *INEX 2007, LNCS 4862*. Springer-Verlag, Berlin/Heidelberg, 2008, pp. 336–347.

[Zir08] Zirin, C., Nastase, V., and Strube, M. (2008). Distinguishing between instances and classes in the Wikipedia taxonomy. In *Proceedings of the 5th European Semantic Web Conference (ESWC 2008)* (Tenerife, Spain, 1–5 June 2008). 2008.

[Zla06] Zlatić, V., Božičević, M., Štefancic, H., and Domazet, M. (2006). Wikipedias: collaborative Web-based encyclopedias as complex networks. Available at: <http://arxiv.org/abs/physics/0602149>. 3 July 2006.

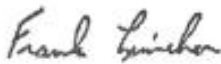
Appendix: IRB Approval Notice



**DREXEL UNIVERSITY
COLLEGE OF MEDICINE**

Office of Regulatory Research Compliance
APPROVAL NOTICE WITH CONSENT

TO: Xia Lin, Ph.D.
Provost / College of Information Sci & Tech
Mailstop: DRE¹

FROM: 
Francis Linnehan, Ph.D., Chair
Institutional Review Board (IRB #3)
Drexel University College of Medicine
1601 Cherry Street, Suite 10444, 3-Parkway, Philadelphia, Pa 19102
Tel: 215-255-7864 Fax: 215-255-7874

SUBJECT: Enabling Entity Retrieval by Exploiting Wikipedia as a Semantic Knowledge Source
SPONSOR: Internal
PROJECT No: 1044303, PROTOCOL No: 19542, ACTION No: 56958 Type: New Period: 1 Seq: 1, DETAIL No: 274001
CURRENT APPROVAL PERIOD: 04/14/2011, EXPIRES: 04/13/2012
USE CONSENT FORM DATED: 04/14/2011

RE: 4/14/11 – According to 45 CFR 46.110, this study is Approved Expedited Category 7. This study will enroll 30 subjects recruited from the General Public.

Included in the Approval: Consent Form and advertisement

Date: 4/26/2011

On behalf of the Committee, I am pleased to inform you that the subject protocol has been reviewed and APPROVED AS SUBMITTED for the period indicated above. We operate under many Government requirements. As a result, this approval is granted with the following understandings:

1. The attached consent form indicated above must be used unless a subsequent notification is approved in writing by the IRB. Remember that each subject enrolled in the study (and/or their guardian) must sign this consent form; preferably, the signatures are witnessed or acknowledged. You must give each subject a copy of the consent form. For record keeping and storage contact the Office of Research Compliance. Please keep these forms readily available (NOT in patients' charts).

1601 Cherry Street, 3 Parkway Building, Suite 10444 • Philadelphia, PA 19102 • Phone 215-255-7857 • Fax 215-255-7874
www.research.drexel.edu • www.drexelmed.edu

In the tradition of Woman's Medical College of Pennsylvania and Hahnemann Medical College®
Philadelphia Health & Education Corporation, a Not-For-Profit Corporation, is the sole proprietor of the Drexel University College of Medicine. Drexel University is not involved in patient care.

2. If this is a sponsored project, then the study may not be activated until the Contract is fully executed by the Clinical Research Group. If this is not a sponsored study (designated "internal"), the costs of the project must be identified and a cost center designated. Please call 215-255-7857 if you have any questions regarding these procedures.
3. You must advise the IRB of the activation date. Use the attached form for this purpose.
4. Any change to the protocol must be submitted in writing and approved by the IRB in advance.
5. Any adverse reaction must be reported to the IRB as soon as it occurs.
6. Should the IRB decide to monitor your project directly, please cooperate fully. Failure to do so may result in withdrawal of this approval and notification to the sponsor and/or Federal agencies. Specific information regarding monitoring appears in **GUIDELINES FOR BIOMEDICAL AND BEHAVIORAL RESEARCH INVOLVING HUMAN SUBJECTS**, and **GUIDELINES FOR NON-MEDICAL** obtainable through this office or the website <http://research.drexel.edu>.
7. Whether or not this protocol is activated, the IRB will conduct Continuing Review at least annually. Should you fail to respond to this Federally-required continuing review and progress report, the project may become ineligible for re-approval and the IRB may choose not to consider other projects for approval.
8. A final progress report must be submitted to the IRB in format similar to that of a periodic report.

The IRB welcomes your research project into the list of approved protocols. Your compliance with the above conditions will help to protect the continuation of all research activity at the University. With your project and others like it, we look forward to additions to knowledge of human health and benefits to science, our patients, and society.

cc: IRB Chair, Dept Chair, Tenet, Drexel

MEMORANDUM
Institutional Review Board (IRB #3)

ACTIVATION NOTICE

TO: Institutional Review Board (IRB #3)
 1601 Cherry Street, Suite 10444, 3-Parkway, Philadelphia, Pa 19102
 Tel: 215-255-7864 Fax: 215-255-7874

FROM: Xia Lin,
 Provost / College of Information Sci & Tech

SUBJECT: ACTIVATION OF HUMAN RESEARCH PROTOCOL ENTITLED:
 Enabling Entity Retrieval by Exploiting Wikipedia as a Semantic Knowledge Source
 PROJECT No: 1044303, PROTOCDL No: 19542, ACTION No: 56958 Type: New Period: 1 Seq: 1,
 DETAIL No: 274001
 DATE OF APPROVAL: 04/14/2011, EXPIRES: 04/13/2012

Date: 4/26/2011

This is to inform the IRB that the subject protocol was activated* on / / . I understand that a Periodic Report for Continuing Review or Final Summary is due on or before the above Expiration Date.

Yes I have a copy of the University's Human Subjects Guidelines and Federal Wide Assurance (FWA) to the OHRP, as required in 45 CFR Part 46.
 No

NOTE:

The University Guidelines for Biomedical and Behavioral Research for the protection of human subjects have been posted on the Office of Research website.

There are two sets of Guidelines - one each for Medical and Non-Medical Research.

You must have a hard copy and read these Guidelines to make sure that these Guidelines are met.

To download a copy of the University Guidelines, follow the below instructions:

1. Go to <http://research.drexel.edu>
2. Click "Medical IRB" or "Non-Medical IRB" in Quick Links
3. Under "Go to", click "Medical IRB" or "Non-Medical IRB Guidelines"
4. Please keep a copy of the University Guidelines in your office.

 (Signed) Lin, Xia

* "Activated" means that the first new human subject was accrued, or an experimental procedure was performed, or records were reviewed under this protocol on or after the date of last approval: 04/14/2011.
Accordingly, this notice must be sent to the IRB ONLY for the FIRST such accrual since that date.

Investigator: Dr. Xia Lin, College of Information Science and Technology, Drexel University

Proposal Title: Enabling Entity Retrieval by Exploiting Wikipedia as a Semantic Knowledge Source

Advertisement to be circulated through the College of IST announcement email list, posted on Drexel University bulletin boards, and verbally introduced at classes at the College of IST:



Would you like to be part of a research study?

Are you interested in test-driving a new information search interface?

If so, you are invited to participate in an experiment for a research project to study information retrieval effectiveness. In the experiment, you will be asked to find answers to questions about movies, movie awards, and actors/actresses. The experiment is a one-time event and will last about one hour and half. You will get paid \$20 for participating. (Any fee you will be paid will be determined by the amount of time you spend in the study.)

To participate in the study you must be between ages 18 and 65 and have experience of searching for information on the Internet using search engines such as Google or on the websites such as Wikipedia.

If you are interested in participating in the study, please contact Sofia at sofia@drexel.edu. Please put "EXPERIMENT" on the email subject line. You can freely decide whether or not to participate, once you get more information.

Study Details:

Title: "Enabling Entity Retrieval by Exploiting Wikipedia as a Semantic Knowledge Source"

Principal Investigator: Dr. Xia Lin, College of Information Science and Technology, Drexel University, 3141 Chestnut Street, Philadelphia, PA 19104

This research is conducted by a researcher who is a member of Drexel University.

APPROVED
Office of Regulatory Research Compliance
Protocol # 19542-01
Approval Date: 4/14/11
Expiration Date: 4/13/12



Drexel University
Consent to Take Part In a Research Study

1. **Subject Name:** _____
2. **Title of Research:** Enabling Entity Retrieval by Exploiting Wikipedia as a Semantic Knowledge Source
3. **Investigator's Name:** Dr. Xia Lin
4. **Research Entity:** Drexel University
5. **Consenting for the Research Study:**
 This is a long and an important document. If you sign it, you will be authorizing Drexel University and its researchers to perform research studies on you. You should take your time and carefully read it. You can also take a copy of this consent form to discuss it with your family member, attorney or any one else you would like before you sign it. Do not sign it unless you are comfortable in participating in this study.
6. **Purpose of Research:**
 You are being asked to participate in a research study. The purpose of this study is to find out if an online information search process can be improved when using an interface that allows the users to specify their queries more directly than by just entering keywords.

This research is being conducted in partial fulfillment of requirement for a doctoral degree.

The study is open to participants who have the ability and experience of doing common search tasks on the Internet using Google, etc. Those who do not have such an ability or experience are excluded from the study. The approximate number of participants to be recruited is 30.

You are free to choose whether or not to participate in the study. You are also free to withdraw from the study at any time you wish.

Your participation in the study may also end in case you do not follow the instructions given or in case some unforeseen circumstances arise which hinder your participation.

APPROVED
 Office of Regulatory Research Compliance
 Protocol # 19542-01
 Approval Date: 4/14/11
 Expiration Date: 4/13/12



Subject Initials _____

7. **PROCEDURES AND DURATION:**

You understand that the following things will be done:

Pre-Experimental Procedures:

You will be introduced to the purpose and general methods of this study. You will then be directed to a webpage which contains background information and detailed instructions for using a new interface. Under the guidance of the study coordinator, you will read through the information and instructions and then test your understanding by trying to find answers for a few sample questions by using the interface. Once you confirm your understanding of how to use the interface, the experimental procedures will start.

Experimental Procedures:

You will be given a question form/sheet which contains 2 sets of 5 questions each. You will be asked to find answers to the two sets of questions by using two different search interfaces. You will be asked to spend no more than 5 minutes per each question. As you find answers for each question, you will be asked to enter the answers on the question form.

Post-Experimental Procedures:

Once you complete the experimental task (or the total time of approximately 50 minutes allocated for completing the task expires), you will be given a questionnaire asking about your experience of doing the task. You will be asked to fill out and submit the questionnaire.

The total duration of the study is expected to be about an hour and half.

Participation in this study is a one-time event. You will not be contacted or asked to participate in a follow-up.

8. **RISKS AND DISCOMFORTS/CONSTRAINTS:**

Possible risks and discomforts you may experience while participating in the study are minor fatigue and boredom, which you may usually experience when searching for information on the Internet under normal circumstances.

9. **UNFORESEEN RISKS:**

Participation in the study may involve unforeseen risks. If unforeseen risks occur, they will be reported to the Office of Regulatory Research Compliance.

10. **BENEFITS:**

There may be no direct benefits from participating in this study. However, the anticipated benefits to the society to be accrued as a result of this study include enhanced understanding of the information search process/mechanism and resultant insights for improvement of information retrieval systems/interfaces for use by the general public.

APPROVED
Office of Regulatory Research Compliance
Protocol # 19942-01
Approval Date: 4/14/11
Expiration Date: 4/13/12



Subject Initials _____

11. **ALTERNATIVE PROCEDURES:**
The alternative is not to participate in this study.
12. **VOLUNTARY PARTICIPATION:**
Participation in this study is voluntary, and you can refuse to be in the study or stop at any time. There will be no negative consequences if you decide not to participate or to stop.
13. **STIPEND/REIMBURSEMENT:**
You will be paid \$20 if you participate in the study. Any fee you will be paid will be determined by the amount of time you spend in the study.
14. **RESPONSIBILITY FOR COST**
Participation in this study will be of no cost to you.
15. **CONFIDENTIALITY:**
In any publication or presentation of research results, your identity will be kept confidential, but there is a possibility that records which identify you may be inspected by authorized individuals, the institutional review boards (IRBs), or employees conducting peer review activities. You consent to such inspections and to the copying of excerpts of your records, if required by any of these representatives.
16. **OTHER CONSIDERATIONS:**
If you wish further information regarding your rights as a research subject or if you have problems with a research-related injury, for medical problems please contact the institution's Office of Regulatory Research Compliance by telephoning 215-255-7857.

APPROVED
Office of Regulatory Research Compliance
Protocol # 15542-01
Approval Date: 4/14/11
Expiration Date: 4/13/12



Subject Initials _____

Version 1

Page 4 of 4

17. CONSENT:

- I have been informed of the reasons for this study.
- I have had the study explained to me.
- I have had all of my questions answered.
- I have carefully read this consent form, have initialed each page, and have received a signed copy.
- I give consent voluntarily.

DO NOT SIGN THIS INFORMED CONSENT AFTER THIS DATE 4/3/12

Subject or Legally Authorized Representative

Date

Investigator or Individual Obtaining this Consent

Date

List of Individuals Authorized to Obtain Consent

Name	Title	Day Phone #	24 Hr Phone #
Sofia Jeon	Research Assistant	215-895-2474	215-895-2474

APPROVED
Office of Regulatory Research Compliance
Protocol# 19542-01
Approval Date: 4/14/11
Expiration Date: 4/13/12



Subject Initials _____

Vita

Sofia Jeon (a.k.a. Sofia J. Athenikos), Ph.D. (, Ph.D.)

Education:

Ph.D. Information Studies	Drexel University, Philadelphia, USA	(September 2011)
M.S. Computer Science	Drexel University, Philadelphia, USA	
Ph.D. Religion	Temple University, Philadelphia, USA	
M.A. Religion	Temple University, Philadelphia, USA	

University Scholarships:

Research Assistantship	College of IST, Drexel University
Research/Teaching Assistantship	Computer Science Department, Drexel University
Teaching Assistantship	Religion Department, Temple University
University Fellowship	Graduate School, Temple University

Selected Honors:

Selection as a recipient of the 2011 Eugene Garfield Doctoral Dissertation Fellowship awarded by Beta Phi Mu (BPM) the International Library & Information Studies Honor Society.

Selection for the ACM Student Research Competition (SRC) at the 20th ACM Conference on Hypertext and Hypermedia (Hypertext 2009) (Torino, Italy, 29 June - 1 July 2009).

Selection as a recipient of the ACM SIGAPP STAP (Student Travel Award Program) Award for paper presentation at the 24th Annual ACM Symposium on Applied Computing (ACM SAC 2009) (Honolulu, Hawaii, USA, 8-12 March 2009).

Selection as a recipient of the Student Poster Award awarded by Google at the Fall 2008 North East Database & Information Retrieval (DB/IR) Day.

Selection as a finalist for the 2008 Google Anita Borg Scholarship.

Induction into Upsilon Pi Epsilon (UPE) the International Honor Society for the Computing and Information Disciplines.

Publications:

Publications in various journals and conference/workshop proceedings, including *Proceedings of the Third International Conference on Design Science Research in Information Systems and Technology (DESRIST 2008)*, *Proceedings of the Chicago Colloquium on Digital Humanities and Computer Science*, *Proceedings of the 2008 International Workshop on Biomedical and Health Informatics (BHI 2008)*, *Proceedings of the 2009 ACM Symposium on Applied Computing*, *Bulletin of IEEE Technical Committee on Digital Libraries (TCDL)*, *Proceedings of the 2009 Symposium on Interactive Visual Information Collections and Activity*, *Digital Humanities 2009 Conference Abstracts*, and *Computer Methods and Programs and Biomedicine*.

Professional Services:

Services as publicity chair, session chair, program committee member, and reviewer for various journals, conferences, and workshops, including ONISW 2007, *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, ER 2008, ACM SAC 2008, ONISW 2008, ACM SAC 2009, SEMAPRO 2009, *ACM SIGMIS DATA BASE for Advances in Information Systems*, ER 2009, ODBASE 2009, *Health Information and Libraries Journal (HILJ)*, ACM SAC 2010, SEMAPRO 2010, *Artificial Intelligence in Medicine (AIIM)*, ACM SAC 2011, WikiSym 2011, SEMAPRO 2011, and ACM SAC 2012.

