Multifaceted Entity/Fact/Relation Retrieval via Semantic Search Interface based on Domain Knowledge Extraction

Sofia J. Athenikos
Drexel University
College of Info Science & Technology
Philadelphia, PA 19104, USA
(1)215-895-2482
sofia.j.athenikos@acm.org

Xia Lin
Drexel University
College of Info Science & Technology
Philadelphia, PA 19104, USA
(1)215-895-2482
linx@drexel.edu

ABSTRACT

A significant kind of information search on the Web is concerned with finding entities of a specific type that satisfy certain semantic conditions. Nevertheless, the conventional keyword-based search mechanism commonly found on the Web does not enable the user to specify the semantic type/conditions for the entities sought, and accordingly does not return the entities as direct search results. The main objective of the project presented in this paper, entitled PanAnthropon FilmWorld, is to demonstrate directly retrieving entities that match the semantic type and conditions specified in the query, by taking a domain-oriented approach to knowledge extraction and retrieval. To this end, the project first constructed a knowledge base containing the semantic information extracted or derived from Wikipedia concerning the film domain. The project then constructed an interactive search interface which provides various semantic search functions besides the main entity retrieval function. The results of evaluation confirm both the effectiveness or semantic information extraction and the effectiveness of direct entity/fact retrieval using the semantic search interface.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *query formulation, retrieval models, search process*; H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *web-based services*; I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods – *relation systems, semantic networks*; I.2.6 [**Artificial Intelligence**]: Learning – *knowledge acquisition*.

General Terms

Design, Experimentation, Human Factors, Performance

Keywords

Entity/Fact/Relation Extraction, Semantic Search, Faceted Search, Entity/Fact/Relation Retrieval, Wikipedia, Semantic Web

WSDM'12, February 8–12, 2012, Seattle, WA, USA.

1. INTRODUCTION

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 keyword-based search interfaces on the Web only allow the user to express information needs in terms of a query string consisting of keywords, and in response return a list of pages that contain all or some of the individual keywords in the query string, rather than a list of the objects of query that directly match the information needs. As such, the matching between the query and the query result does not take semantics into account.

Wikipedia (http://www.wikipedia.org/) has become an important semantic knowledge source, due to its 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, which is connected to other entities via explicit or implicit semantic relations.

The research problem addressed by the project presented in this paper, entitled PanAnthropon FilmWorld, is how to effectively enable and facilitate 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.

The PanAnthropon project addressed the problem by exploiting Wikipedia as a semantic knowledge source, with the film domain as its initial proof-of-concept domain of application. By building a semantic knowledge base containing domain-relevant classes, entities, attributes, and facts extracted or derived from Wikipedia and by implementing and evaluating a semantic search interface connected to the knowledge base, the project has demonstrated the utility and feasibility of retrieval of entities and related facts that directly match the user's information needs.

The rest of this paper is organized as follows: Section 2 discusses related work. Section 3 outlines the conceptual basis of this work. Section 4 briefly describes the process and results of semantic information extraction. Section 5 illustrates the query functions enabled by the semantic search interface. Section 6 discusses the method and results of the evaluation on the effectiveness of information extraction. Section 7 discusses the method and results of the evaluation on the effectiveness of information retrieval using the search interface. Section 8 concludes the paper.

2. RELATED WORK

The related work can be discussed in terms of three relevant areas: (1) entity search, retrieval, and ranking; (2) info extraction from Wikipedia; and (3) info retrieval based on Wikipedia data.

Entity search/retrieval/ranking is an emerging field of information retrieval that aims to retrieve/rank entities that match a given query. This project is concerned with entity search/retrieval, not with ranking, because it considers only *exact semantic matching*.

The problem of finding and ranking entities on the Web has been studied by Cheng et al. [7,8]. The problem of ranking (related) entities identified from the documents returned by a standard search engine has been investigated by Zaragoza et al. [14]. The task of retrieving and ranking semantic-type-specified entities that match narrative descriptions given in the queries has been taken by the INitiative for Evaluation of XML Retrieval (INEX), which in 2007 started the Entity Ranking Track [9], and by the Text REtrieval Conference (TREC), which in 2009 initiated the Entity Track (http://trec.nist.gov/data/entity09.html). Nevertheless, all these approaches more or less still operate within the word-based, document-centric framework of traditional information retrieval. This project stands out in that its entity-centric, semantics-based approach encompasses the process from entity extraction and indexing to entity search and retrieval.

The task of extracting large-scale semi-structured data from Wikipedia has been attempted by Suchanek et al. [12, 13] and Auer et al. [3,4], with the YAGO (Yet Another Great Ontology) project (http://www.mpi-inf.mpg.de/yago-naga/yago/index.html), and the DBpedia project (http://www.dbpedia.org/), respectively.

YAGO is built upon a data model within the framework of description logics [5]. Even though this project did not attempt at providing a model-theoretic framework for defining the semantics of the data elements and their relations, the data model underlying this project is similar to the one in YAGO. In contrast to YAGO, which is concerned with general-domain or multi-domain info extraction/retrieval, this project focuses on domain-oriented info extraction/retrieval. Even though, accordingly, this project used a relatively small, selected subset of Wikipedia, a comparison of the number of entities and facts extracted, with respect to the relative sizes of the source datasets, shows that on average this project extracted or derived more entities and facts per Wikipedia page. While this project processed structured templates as in the YAGO project, it also processed, in a knowledge-aware manner, some unstructured portions of Wikipedia pages, which are far more difficult to process to extract semantic info with high accuracy. The DBpedia project, as another Semantic Web [6] project, is similar to YAGO in its intents and purposes. In general, similar remarks can be made in comparing DBpedia with this project.

LinkedMDB (Linked Movie Database) (http://linkedmdb.org/) [10], which mainly used Freebase (http://www.freebase.com/) as the source for info extraction, is related to this project in terms of application domain, even though it did not extract info directly from Wikipedia. Again, a comparison shows that on average this project extracted more entities per film than LinkedMDB.

Although some research-based prototype search systems exist, e.g., Koru by Milne et al. [11], which use the lexical information extracted from Wikipedia so as to facilitate keyword-based page retrieval, the examples of search systems that use the semantic knowledge extracted from Wikipedia for the task of retrieving entities are provided, again, by YAGO and DBpedia. Both DBpedia and YAGO provide interfaces for querying the semantic knowledge extracted from Wikipedia by using SPARQL (http://www.w3.org/TR/rdf-sparql-query/) patterns composed of a set of conditions, each of the form <subject, predicate, object>.

The YAGO query form provides a dropdown menu containing all available predicates to choose from. Since YAGO is a general-or multi-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 menu consists of all predicates regardless of whether or not a given predicate may be applicable to the subject entity. The DBpedia query form and query input format are quite similar to those of the YAGO interface, except the fact that the predicate field here provides suggestions using the look-ahead technology, and except the fact that query results are presented in a table format rather than in a list format.

The interface constructed from this project is similar to those of YAGO and DBpedia in appearance. However, it provides multiple types of semantic search/retrieval functions, which are facilitated by explicit specification of entity type/subtype and by interactive menu option presentation reflecting the conceptual framework.

3. CONCEPTUAL FRAMEWORK

In this project, “entities” are conceived of as things of all kinds that can be classified into different “classes” and that have certain “attributes”. The kinds of classes and attributes that are relevant depend on the domain at issue (i.e., the ontological space). This project therefore takes a domain-oriented approach to ontology construction as well as knowledge extraction and retrieval.

The film-domain-oriented ontology constructed from project is at: http://dlib.ischool.drexel.edu:8080/sofia/PA/Ontology.pdf. Each column in the ontology table corresponds to a distinct level in the subsumption hierarchy, from the top level to level 5. The entities extracted/derived in this project are semantically typed according to this ontology. Specifically, the “type” of an entity refers to the level-1 class, while the “subtype” refers to the leaf class subsumed by the former. A simplified entity classification scheme is at: http://dlib.ischool.drexel.edu:8080/sofia/PA/Ontology_Simple.pdf The simplified scheme is used for the entity type/subtype menu presentation on the search interface.

As suggested above, different attributes apply to different entities, depending on their types/subtypes. As in the case of classes, new attributes were extracted or derived, according to the progress of direct extraction and indirect derivation of semantic knowledge. A table containing the list of 190 attributes, along with information on the applicable types of entities, values, and value entities, is at: http://dlib.ischool.drexel.edu:8080/sofia/PA/Attributes.html.

Another classification scheme was also built and used to classify the Wikipedia categories extracted in a systematic manner. The taxonomy consisting of 215 super-categories is partially shown at: http://dlib.ischool.drexel.edu:8080/sofia/PA/SuperCategories.pdf. Unlike in Wikipedia, only one leaf super-category was assigned to a given regular category.

In this project, a “fact” concerning an entity refers to a tuple in the form of <entity, attribute, value, note>, where “value” can consist of a literal, an entity, a class, or a Wikipedia category. (When another entity occupies the value position, it represents a relation between the two entities.) The “note” field is used to store contextual information relevant to a given fact, which is not possible in a strict <subject, predicate, object> model.

4. SEMANTIC INFO EXTRACTION

The extraction system was implemented by using Java servlets, Tomcat server, and MySQL database. The details concerning the process of direct extraction and indirect derivation of semantic information using Wikipedia are presented in another paper [1]. Here only a brief description of the process is given, followed by the description of the storage/organization of the extracted/derived information and the statistics on the extraction/derivation results.

4.1 Info Extraction/Derivation Process

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. For this purpose, Wikipedia category page “Years in film” (http://en.wikipedia.org/wiki/Category:Years_in_film) was used to extract the titles/URLs of 120 pages corresponding to each year in film history between years 1890 and 2009, inclusive. The 120 pages were subsequently downloaded.

A total of 11,355 film titles were extracted from each page in the 120-page set. Each film in the 11,355-film set was considered as an entity and was entered into a database table. Wikipedia pages for 10,640 films that have corresponding articles were then downloaded and served as the main source of information.

For efficient processing, relevant sections of the downloaded film pages — abstract, infobox, categories, and film cast info section — were separately stored for each film in a database table, and information extraction was done by retrieving and processing each section separately, for all films at once, in turn.

The abstract section was mainly used to extract “also_known_as” facts on a given film and a brief introductory info excerpt to be provided for the film via the Slide function of the search interface. The infobox section was used to extract film-relevant attributes, e.g., “directed_by”, “produced_by”, etc., and corresponding facts and associated entities that serve as the values for those facts. The categories section was used to extract categories associated with a film page and corresponding “associated_with_category” facts. The film cast info section was used to extract “has_cast_member” facts, which relate a given film with its cast members, and associated person entities that serve as the values for those facts. The role(s) played by a cast member, if any, were stored in the note field used to store context information for a given fact.

In addition, Wikipedia pages about two well-known film awards, i.e., Academy Awards and Golden Globe Awards, were also downloaded and processed in order to extract facts on the award events and award winners/nominees of selected award categories for each year (up to year 2010) of the award ceremonies.

The facts directly extracted by processing Wikipedia pages were used to derive more classes, entities, attributes, and facts.

Upon completion of indirect information derivation, all unique attributes were entered into a database table with info on the applicable types of entities, values, and value entities. Then all entity-centric facts were transformed into attribute-centric facts and saved in a separate table with the additional information.

4.2 Info Organization/Storage

The semantic information extracted/derived from Wikipedia has been stored in the MySQL database by using three data models: (1) Hierarchical Tree Model; (2) Common Relational Model; and (3) Entity–Attribute–Value Model.

Table 1 presents the three data models and corresponding data representation formats. (The Entity–Attribute–Value Model has three variations of data representation, depending on the types of facts represented, i.e., entity-centric, category-centric, or attribute-centric.) Table 2 shows the grouping of database tables according to underlying data models.

Table 1. Data models and data representation formats

Data Model	Data Representation Format
Hierarchical Tree Model	<child_node, parent_node>
Common Relational Model	<element, field_1, field_2, ..., field_n>
Entity–Attribute–Value Model	<entity, attribute, value, note> <category, attribute, value, note> <attribute, entity, value, note>

Table 2. Database tables grouped by data models

Data Model	Database Table	Table Content
Hierarchical Tree Model	Class	classes
	Category_Super	super-categories
Common Relational Model	Entity	entities
	Category	categories
	Attribute	attributes
	Page	film entity pages
Entity–Attribute–Value Model	Entity_Fact	entity-centric facts
	Category_Fact	category-centric facts
	Attribute_Fact	attribute-centric facts

4.3 Info Extraction/Derivation Statistics

Table 3 presents the extraction/derivation 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 4 shows the number of entities per entity type.

Table 3. Overall information extraction/derivation statistics

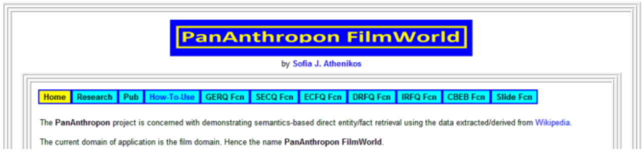
Database Table	Record Type	Count
Class	class	72
Page	film entity page	10,640
Entity	entity	209,266
Entity_Fact	entity-centric fact	2,354,931
Category_Super	super-category	215
Category	category	5,229
Category_Fact	category-centric fact	91,335
Attribute	attribute	190
Attribute_Fact	attribute-centric fact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

Table 4. 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

5. SEMANTIC SEARCH INTERFACE

The Web interface for the PanAnthropon FilmWorld project was implemented by using HTML, JavaScript, and JSP (in connection with the MySQL database) on the Tomcat server. The interface (Figure 1) is at: http://dlib.ischool.drexel.edu:8080/sofia/PA/. A brief description of the design of the interface is provided in [2]. Here the search functions offered by the interface (excluding the Slide function that presents the image and introduction for each film via a single menu containing the list of films) are described and illustrated in greater detail, with a focus on the main function.



The **PanAnthropon** project is concerned with demonstrating semantics-based direct entity/fact retrieval using the data extracted/derived from Wikipedia.

The current domain of application is the film domain. Hence the name **PanAnthropon FilmWorld**.

Figure 1. PanAnthropon FilmWorld interface.

5.1 General Entity Retrieval Query

The General Entity Retrieval Query (GERQ) function is one that corresponds to the main research problem of this project, namely, to demonstrate retrieving entities that directly match a given query that specifies the type/subtype and conditions (i.e., <attribute, value> pairs) to be satisfied by the entities. Figure 2 shows the flowchart of GERQ input process. (In the diagram, the double-arrow connector (↔) represents the steps that can be repeated.)

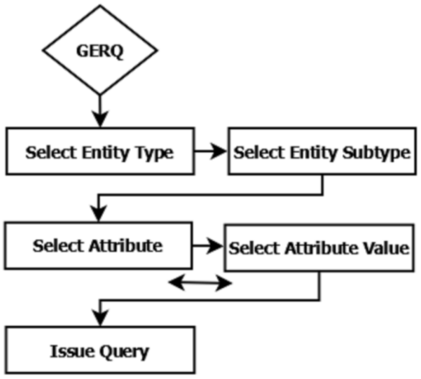
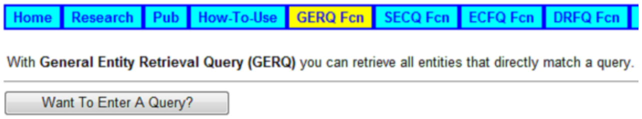


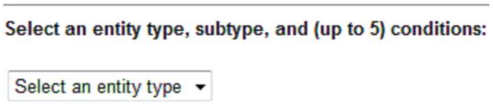
Figure 2. GERQ query process flowchart.

Figure 3 shows the initial screen of the GERQ interface. Once the user clicks on the "Want To Enter A Query?" button, the query form appears, which, at this stage, contains only a menu for entity type selection, as shown in Figure 4. Once the user selects an entity type, a menu for entity subtype selection appears, as shown in Figure 5. The menu contains only those entity subtypes that are relevant to the selected entity type. In the case of entity types "person" and "organization", a simplified menu appears, as shown in Figure 6, which does not differentiate between entity subtypes.



With **General Entity Retrieval Query (GERQ)** you can retrieve all entities that directly match a query.

Figure 3. GERQ interface initial screen.



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

Figure 4. GERQ initial query form.

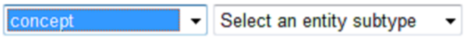


Figure 5. GERQ query form after entity type selection.

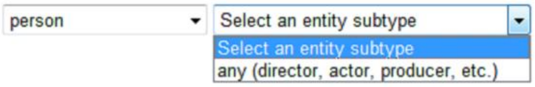


Figure 6. GERQ simplified entity subtype selection menu.

Once the user selects an entity subtype, a menu for attribute selection appears, as shown in Figure 7. The menu contains only relevant attributes to choose from, according to the entity type and subtype selected.

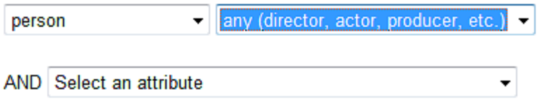


Figure 7. GERQ query form after entity subtype selection.

Once the user selects an attribute, an input box may appear, as shown in Figure 8, so that the user can start typing to get a menu of suggested values. 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 9. Once the user selects a value for the selected attribute, buttons appear at the bottom of the query form, as shown in Figure 10. The user can submit the query as is, add another condition, or remove the last condition, by clicking on an appropriate button. Once the user submits the query, query processing starts.

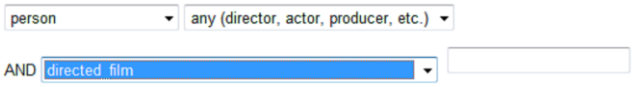


Figure 8. GERQ query form after attribute selection #1.

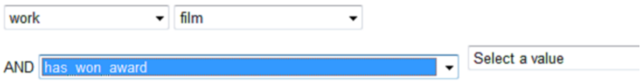


Figure 9. GERQ query form after attribute selection #2.

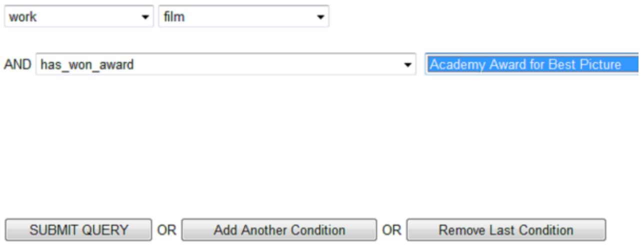
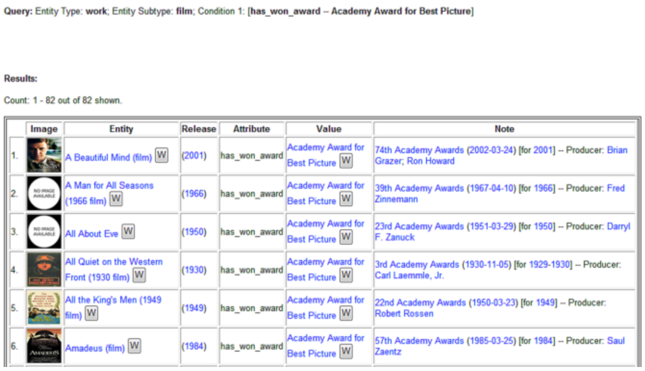


Figure 10. GERQ query form after value selection.

Figure 11 presents a partial snapshot of the result of a sample GERQ query. As shown, the query 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.) 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 the Specific Entity-Centered Query function) appears, as shown in Figure 12. If the user clicks on the "W" button that appears after the name of an entity, a separate window for the corresponding Wikipedia page appears, as shown in Figure 13.



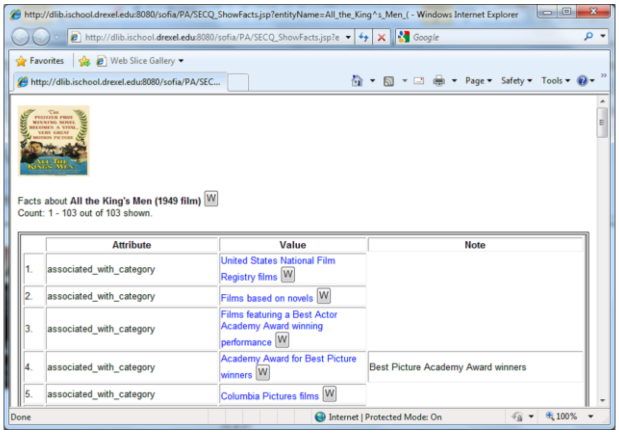
Query: Entity Type: **work**; Entity Subtype: **film**; Condition 1: **[has_won_award -- Academy Award for Best Picture]**

Results:

Count: 1 - 82 out of 82 shown.

	Image	Entity	Release	Attribute	Value	Note
1.		A Beautiful Mind (film) W	(2001)	has_won_award	Academy Award for Best Picture W	74th Academy Awards (2002-03-24) [for 2001] -- Producer: Brian Grazer; Ron Howard
2.	NO IMAGE AVAILABLE	A Man for All Seasons (1966 film) W	(1966)	has_won_award	Academy Award for Best Picture W	39th Academy Awards (1967-04-10) [for 1966] -- Producer: Fred Zinnemann
3.	NO IMAGE AVAILABLE	All About Eve W	(1950)	has_won_award	Academy Award for Best Picture W	23rd Academy Awards (1951-03-29) [for 1950] -- Producer: Darryl F. Zanuck
4.		All Quiet on the Western Front (1930 film) W	(1930)	has_won_award	Academy Award for Best Picture W	3rd Academy Awards (1930-11-05) [for 1929-1930] -- Producer: Carl Laemmle, Jr.
5.		All the King's Men (1949 film) W	(1949)	has_won_award	Academy Award for Best Picture W	22nd Academy Awards (1950-03-23) [for 1949] -- Producer: Robert Rossen
6.		Amadeus (film) W	(1984)	has_won_award	Academy Award for Best Picture W	57th Academy Awards (1985-03-25) [for 1984] -- Producer: Saul Zaentz

Figure 11. GERQ query result.



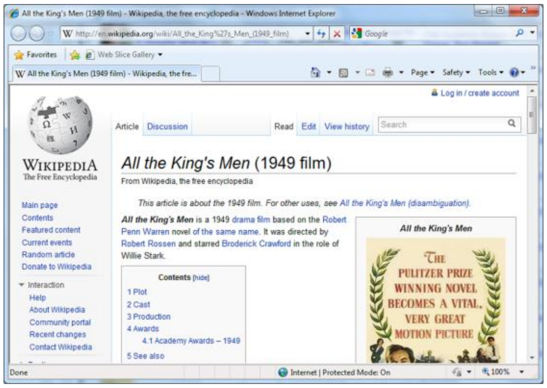
http://dlib.ischool.drexel.edu:8080/sofia/PA/SECQ_ShowFacts.jsp?entityName=All_the_King^s_Men_(- Windows Internet Explorer

http://dlib.ischool.drexel.edu:8080/sofia/PA/SECQ_ShowFacts.jsp?e

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

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

Figure 12. GERQ entity fact window.



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

http://en.wikipedia.org/wiki/All_the_King%27s_Men_(1949_film)

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]
1 Plot
2 Cast
3 Production
4 Awards
4.1 Academy Awards – 1949
5 See also

All the King's Men

Figure 13. GERQ entity Wikipedia page window.

5.2 Specific Entity-Centered Query

The Specific Entity-Centered Query (SECQ) function enables the user to retrieve all entity-centric facts, given the type, subtype, and name of a specific entity. (It can thus be alternatively named Specific Entity Fact Query (SEFQ) or Entity Fact Retrieval Query (EFRQ).) Figure 14 shows the flowchart of SECQ input process.

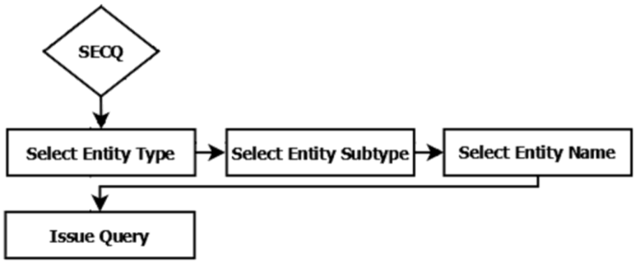
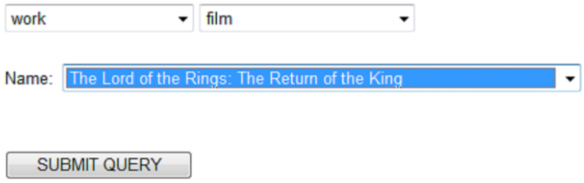


Figure 14. SECQ query process flowchart.

As in GERQ, the initial query form for SECQ only contains a menu for entity type selection. Upon user selection of an entity type, a menu for entity subtype selection appears. Once the user selects an entity subtype, either an input box appears so that the user can get a menu of suggested values or a menu for value selection immediately appears. Once the user selects the name of the entity, as shown in Figure 15, query processing starts. The result of a SECQ query looks similar to Figure 12, which presents the entity-centric facts in the form of <attribute, value, note>.



work ▾ film ▾

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

Figure 15. SECQ query form after entity name selection.

5.3 Entity Commonality Finder Query

The Entity Commonality Finder Query (ECFQ) function refers to retrieving commonalities between two specified entities of the identical entity type and subtype. Here commonalities mean commonly-shared <attribute, value> pairs. Figure 16 shows the flowchart of ECFQ input process. The menu presentation at each step of the process is done interactively as in GERQ and SECQ.

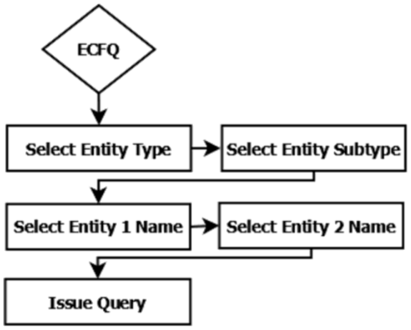
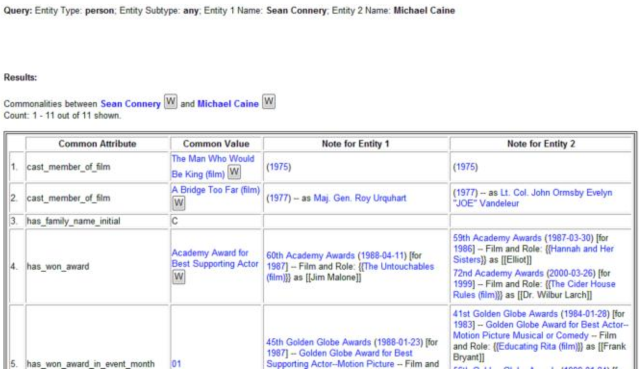


Figure 16. ECFQ query process flowchart.

Figure 17 shows a partial snapshot of the result of a sample ECFQ query. As shown, the result is presented in the form of <attribute, value, note_1, note_2>, where attribute and value represent an attribute–value pair commonly shared by entity 1 and entity 2, and note_1 and note_2 provide respective contexts for entity 1 and entity 2 for the 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 W and Michael Caine W
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) W	(1975)	(1975)
2.	cast_member_of_film	A Bridge Too Far (film) W	(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 W	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 17. ECFQ query result.

5.4 Direct Relation Finder Query

The Direct Relation Finder Query (DRFQ) function allows the user to retrieve direct relations between two specified entities, regardless of their entity types and subtypes. Figure 18 shows the flowchart of DRFQ input process. Figure 19 presents a partial snapshot of the result of a sample DRFQ query. As shown, the result consists of <entity_1, relation, entity_2, note> tuples, which represent <entity, attribute, value, note> tuples where entity 1 and entity 2 occupy the position of entity and of value, respectively.

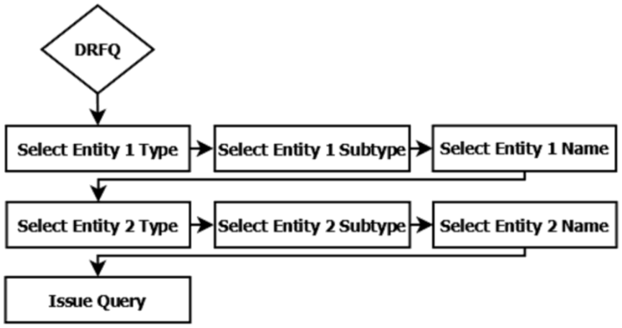
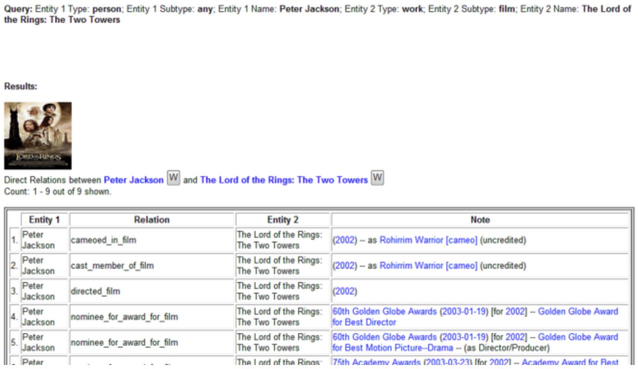


Figure 18. DRFQ query process flowchart.



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 W and The Lord of the Rings: The Two Towers W
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)
	Peter		The Lord of the Rings:	75th Academy Awards (2003-03-23) [for 2002] -- Academy Award for Best

Figure 19. DRFQ query result.

5.5 Indirect Relation Finder Query

The Indirect Relation Finder Query (IRFQ) function enables the user to retrieve 1-degree indirect relations between two specified entities. As shown in Figure 20, the IRFQ query input process is the same as that of DRFQ. As shown in Figure 21, the result of a query using IRFQ consists of <e1, e1-e3_rel, e3, e3-e2_rel, e2> tuples, where e1 and e2 stand for the two specified entities, e3 stands for a third, intermediary entity, and el-e3_rel and e3-e2_rel stand for the relation between entity 1 and entity 3 and between entity 3 and entity 2, respectively.

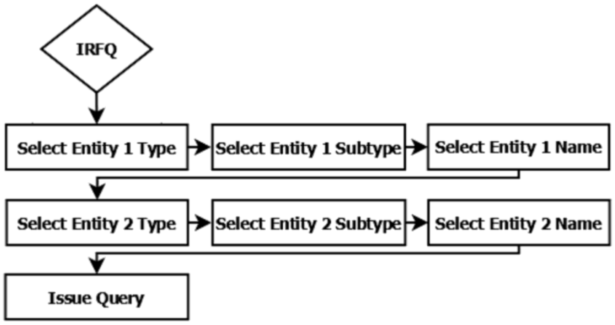
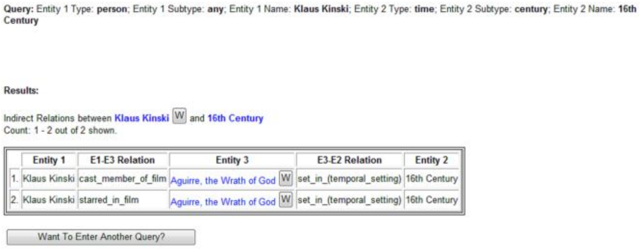


Figure 20. IRFQ query process flowchart.



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 W 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 W	set_in_(temporal_setting)	16th Century
2.	Klaus Kinski	starred_in_film	Aguirre, the Wrath of God W	set_in_(temporal_setting)	16th Century

Figure 21. IRFQ query result.

5.6 Category-Based Entity Browsing

The Category-Based Entity Browsing (CBEB) function refers to retrieving (only) film entities by using the taxonomy of super-categories and categories. Figure 22 shows the flowchart of CBEB input process. Once the user selects a top-level super-category, menus for sub-super-category selection progressively appear, until a menu for leaf-level category selection appears. Once the user selects a leaf category, query processing starts. The query result presents the image, title, release year, and Wikipedia page button for each film that has been directly assigned the selected category.

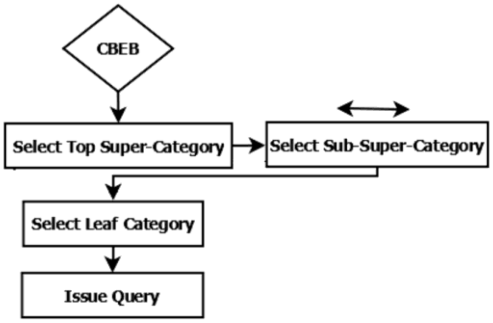


Figure 22. CBEB query process flowchart.

6. INFO EXTRACTION EVALUATION

The first evaluation has been performed in order to validate the quality of data extracted by the information extraction system of this project, compared against the source data in Wikipedia. The quality of data is evaluated in terms of two criteria: (1) Precision: How much of the extracted/derived data is accurate? (2) Recall: How much of the data in the source has been extracted/derived? The two criteria are measured by using the equations shown in Figure 23, which are analogous to the equations to compute precision and recall in conventional information retrieval.

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

Figure 23. Equations for IE evaluation.

6.1 Evaluation Dataset

Given the fact that the main source of info extraction/derivation in this project consisted of 10,640 Wikipedia pages on films, the test set for evaluation was constructed by retrieving the data extracted or derived from 100 film pages. In order to evaluate data quality in a balanced manner, 50 films 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) involving each film in the 100-film set were retrieved to be compared with the source Wikipedia pages.

6.2 Evaluation Methods

The evaluation was performed by manually inspecting 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 info sections of the 100 source Wikipedia pages.

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 as correct. If no such fact was extracted, then it was considered as a missing fact. If, on the other hand, the source page does not have a category link as mentioned above but a fact as described above was nevertheless extracted, then the extracted fact was considered as an incorrect fact. The facts that were indirectly derived from the categories, based on the taxonomy of super-categories, e.g., <film, belongs_to_genre_of, Epic Film (genre)>, were also inspected to determine correctness.

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 about directors, producers, writers, narrators, starring actors, cinematographers, editors, musicians, studios, distributors, release dates, running times, countries, and languages associated with a film.

The facts extracted from the abstract include also_known_as facts and the facts about directors, producers, writers, and starring actors (in case such info was not given in the infobox). Both types of facts were checked against the source to determine correctness.

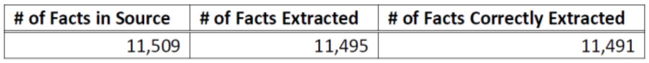
The facts extracted/derived from the film cast info section include has_cast_member facts and has_role facts. If the section contains information about "Clark Gable" being a cast member as "Rhett Butler", then the fact <film, has_cast_member, Clark Gable, as Rhett Butler> was considered correct. Accordingly, 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 info on the roles played by cast members, the correctness of facts extracted was judged by considering only the actor names.

Based on the results of inspection as described above, precision and recall scores (in percentage) were first computed for each film individually, and average precision and recall scores were then computed for all films as a whole. (For the sake of computation of precision/recall scores, the (correct) facts that have been indirectly derived 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.

6.3 Evaluation Results

Table 5 shows the result of evaluation in terms of the number of facts explicitly/implicitly in source vs. number of facts extracted or derived vs. number of facts correctly extracted or derived.

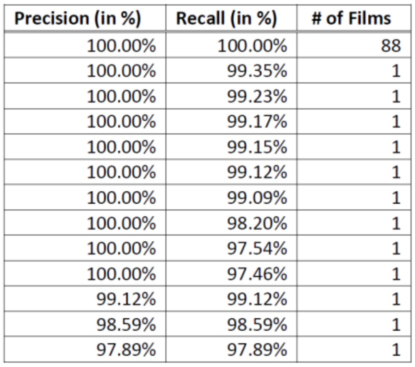
Table 5. IE evaluation result: number of facts



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

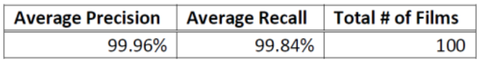
Table 6 shows the number of films in the test set for each distinct precision/recall score pair. It shows that per-film precision/recall scores for 88 out of 100 films in the set were 100% precision and 100% recall. Table 7 shows the average precision/recall scores for the test set as a whole. As shown, the result confirms high data quality with 99.96% average precision and 99.84% average recall.

Table 6. IE evaluation result: 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 7. IE evaluation result: average precision/recall



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

7. INFO RETRIEVAL EVALUATION

The second evaluation has been performed in order to gauge the effectiveness of information retrieval using the search interface constructed from this project. The purpose was to show that the mechanism of retrieving entities and facts by type-and-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. The main intent of this evaluation was not to demonstrate the usability or user-friendliness of the interface, interpreted as ease or simplicity of use, but to demonstrate the effectiveness of info retrieval using the interface. The effectiveness of info retrieval is measured in terms of precision and recall, computed by using the equations shown in Figure 24, which are analogous to the equations used for the evaluation on info extraction.

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

Figure 24. Equations for IR evaluation.

7.1 Experimental Design

Although Wikipedia served as the source of info extracted/derived through this project, it was decided not to use the Wikipedia interface in the evaluation on info retrieval. The decision was based on the consideration for fairness, given that Wikipedia is a general- or multi-domain information source and that, as such, the search result returned when using Wikipedia may include items that are not relevant to the film domain. (The decision not to use YAGO or DBpedia was also based on similar reasons.) Instead, the Internet Movie Database (IMDb) (http://www.imdb.com/) site was chosen for comparison, given 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 evaluation was performed by conducting an experiment with human subjects that represent potential users. The main task of the experiment 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 testing both interfaces, instead of using two distinct groups of subjects to try 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.

The hypotheses tested through the experiment are as follows:

- H1: Per-subject average precision/recall will generally be higher for the PanAnthropon subset (i.e., the subset of questions that the given subject answered by using the PanAnthropon interface) than for the IMDb subset.
- H2: Per-group average precision/recall will be higher for the PanAnthropon subset than for the IMDb subset.

7.2 Experimental Procedures

A total of 33 voluntary subjects were recruited to participate in the experiment. 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. The procedures used for each experimental session (except signing of the informed consent form and the compensation receipt) are described below.

7.2.1 Pre-Task Procedures

After being briefed on the purpose and methods of the study, the subjects were shown "How-To-Use" page of the PanAnthropon site (http://dlib.ischool.drexel.edu:8080/sofia/PA/UserInfo.html). The subjects were asked to read the general background info 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 try on the GERQ interface to see if they could find answers. (The subjects were given only about 5 minutes in total to read the info/instructions and try sample queries.) The subjects were then directed to the IMDb homepage. Most subjects except only a few were quite familiar with the IMDb interface and did not require any practice. 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. 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-task questionnaire consisting of 6 questions. The pre-task procedures were completed with the subjects filling out the questionnaire.

7.2.2 Main Task Procedures

All subjects were administered the same task set consisting of 10 questions, divided into two subsets of 5 questions. One half of the subjects (*N*=12) answered Subset 1 using IMDb and then Subset 2 using PanAnthropon (PA); 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 later.) (The instructions, "Use IMDb Interface" or "Use PanAnthropon Interface", were given before each subset on the task sheets.) Three variations of question ordering were used for each question subset, as shown in Table 8. (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 8. IR evaluation main task design per task code

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

7.2.3 Post-Task Procedures

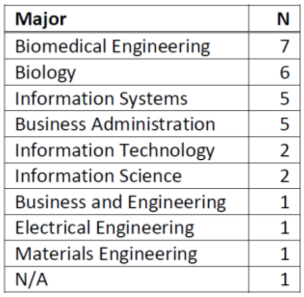
Once the subjects completed the main experimental task, they were given a post-task questionnaire consisting of 8 questions. The experimental session was concluded with the subjects filling out the questionnaire.

7.3 Experimental Results

7.3.1 Pre-Task Questionnaire Responses

The pre-task questionnaire consisted of questions used to gather demographic data on the subjects. Out of 31 subjects to whom the questionnaire was administered, 30 subjects identified themselves as students (undergraduate: *N*=27; graduate: *N*=3). Table 9 shows the major fields of study given by the subjects. The average age of the subjects was 21 (min age = 18; max age = 45). In response to a question on the online info search experience level, 13 subjects marked “Expert”, 17 subjects selected “Intermediate”, and only 1 subject recorded “Novice”. 30 out of 31 subjects indicated that they engage in online search activities several times per day; only 1 subject indicated the lesser frequency of about once per day.

Table 9. IR evaluation pre-task result: subject 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

7.3.2 Main Task Results

The main task set consisted of questions asking for films, people, and film award events, such as:

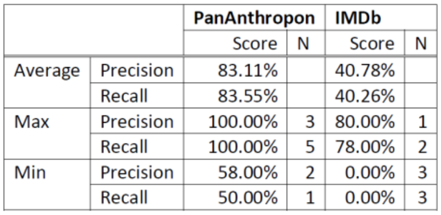
- Who played all of these roles: Clem, Fox, and Hickey?
- Which films produced winner of Academy Award for Best Director and nominee for Academy Award for Best Actor?
- At which (Academy or Golden Globe) award events did Peter Ustinov win awards?
- Which film was directed by Werner Herzog, and has Klaus Kinski as a cast member, and belongs to the genre of Adventure Drama Film, and is set in the 16th century, and has a role named Don Fernando de Guzman?

All the questions in the task set were formulated to have definitely correct answers. Some questions involve sub-questions (omitted above) that ask for query-relevant entities/facts related to the entities returned as the main result. To be considered completely correct, the answers to such questions must include the additional information requested. On the other hand, if an item in the answer to such a question contains all or some of additional information requested but does not contain the correct main entity name, then such an answer item is considered completely wrong. Therefore, a weighted correctness scoring scheme was used for each question.

The analysis of main task results involved computing precision and recall for each test question per subject, computing average precision/recall for each question subset per subject, computing average precision/recall for the subject group as a whole, and analyzing the results in terms of the comparison between IMDb and PanAnthropon. (Due to various problems encountered during experimental sessions, main task data from 9 subjects have been excluded from analysis in order to ensure a valid assessment.)

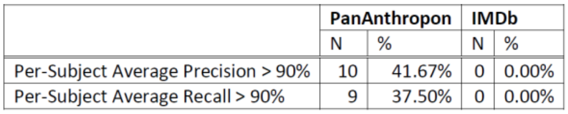
Tables 10–12 show the results of the analysis of 24 subjects’ task results. As shown, the subjects’ info retrieval task performance on the PanAnthropon interface clearly surpassed their performance on the IMDb interface, confirming both H1 and H2, despite the fact that the subjects had an extremely limited exposure to the PanAnthropon interface prior to performing the task.

Table 10. IR evaluation task result: per-group average/max/min precision/recall



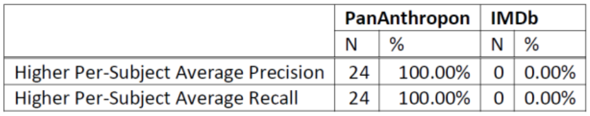
		PanAnthropon		IMDb	
		Score	N	Score	N
Average	Precision	83.11%		40.78%	
	Recall	83.55%		40.26%	
Max	Precision	100.00%	3	80.00%	1
	Recall	100.00%	5	78.00%	2
Min	Precision	58.00%	2	0.00%	3
	Recall	50.00%	1	0.00%	3

Table 11. IR evaluation task result: number of subjects w/ average precision/recall > 90%



	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 12. IR evaluation task result: number of subjects w/ higher precision/recall on PA

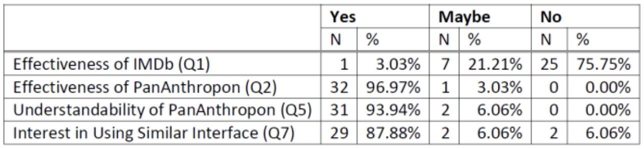


	PanAnthropon		IMDb	
	N	%	N	%
Higher Per-Subject Average Precision	24	100.00%	0	0.00%
Higher Per-Subject Average Recall	24	100.00%	0	0.00%

7.3.3 Post-Task Questionnaire Responses

All 33 subjects’ post-task questionnaire responses have been collected and analyzed. Table 13 shows the summary result on Yes/May/No questions. As shown, 32 subjects answered “Yes” on the effectiveness of the PanAnthropon interface. Despite the fact that the subjects were introduced to the new, unfamiliar interface for the first time and that they were given a very limited amount of practice time, 31 subjects answered “Yes” on the usability and understandability of the PanAnthropon interface. Furthermore, 29 subjects indicated that they would be interested in using interfaces similar to PanAnthropon for info retrieval tasks. Albeit not shown in Table 13, all 33 subjects unanimously agreed on the superior effectiveness of PanAnthropon in contrast to IMDb. The reasons for its effectiveness, given by the subjects, included: (a) no need to guess right keywords; (b) step-by-step search process; (c) ease of finding and selecting applicable entity types/subtypes and attributes; (d) ability to search for specific entities; (e) ability to specify multiple conditions; (f) no need to browse multiple pages to find answers; (g) ease of making comparisons; (h) absence of extraneous information in query results.

Table 13. IR evaluation post-task result: yes/maybe/no



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

8. CONCLUSION

The PanAnthropon FilmWorld project set out to demonstrate the utility, feasibility, and effectiveness of entity-centered, semantic-knowledge-based, and domain-oriented information retrieval. In particular, the project aimed at enabling and validating an entity search mechanism that allows the user to explicitly specify the semantic type/subtype and conditions and to directly retrieve the entities sought (and relevant facts) as the direct results of search.

With the film domain as an initial proof-of-concept domain of application, and with Wikipedia as a semantic knowledge source, the project approached the task by constructing a knowledge base using the semantic information directly extracted and indirectly derived from Wikipedia and by implementing a semantic search interface with the proposed retrieval capability.

In contrast to comparable projects such as YAGO and DBpedia, this project focused on domain-oriented information extraction and retrieval. In extracting semantic information, this project did not only use structured templates and category structures but also utilized unstructured or non-standardized portions of Wikipedia pages. This project also provides an interactive search interface that allows the user to query the content of the knowledge base in an intuitive, step-by-step manner via multiple types of semantic search functions that focus on different semantic facets.

The evaluation on the effectiveness of information extraction, performed via manual inspection of 11,495 film-centric facts extracted/derived concerning 100 films, has confirmed high data quality with 99.96% average precision and 99.84% average recall. The evaluation on the effectiveness of information retrieval using the search interface, performed via experiment using 33 human subjects, has confirmed not only the high utility and effectiveness of the interface but also the high usability and understandability of the interface. Given that the subject population represents typical information seekers who frequently engage in online information search activities, it can be safely reasoned that the results are applicable to a broader population.

The main contribution from this project therefore consists in showing the utility, feasibility, and effectiveness of entity/fact retrieval, successfully achieving the main goal set for this project. Additional contributions include the dataset and the interface themselves as resources that can be utilized beyond this project. The interface is already publicly accessible. Part of the dataset consisting of entities and entity-centric facts, converted to XML, will soon be made freely accessible for research purposes.

The approach used in this project can be applied to domains other than the film domain and to other semi-structured data sources on the Web besides Wikipedia, with some modification. Therefore, the results from this project have far-reaching implications.

9. ACKNOWLEDGMENTS

The research work presented in this paper has been partially supported by the 2011 Eugene Garfield Doctoral Dissertation Fellowship awarded to the first author by Beta Phi Mu the International Library & Information Studies Honor Society.

10. REFERENCES

[1] Athenikos, S.J, and Lin, X. Enabling type/condition-specified entity/fact retrieval using semantic knowledge extracted from Wikipedia. (To be presented at the First International Workshop on Search & Mining Entity-Relationship Data (SMER'11) (Glasgow, UK, 28 October 2011), co-located with the 20th ACM Conference on Information and Knowledge Management (CIKM 2011).)

[2] Athenikos, S.J., and Lin, X. Search as you think and think as you search: semantic search interface for entity/fact retrieval. (Submitted to the Fifth Workshop on Human-Computer Interaction and Information Retrieval (HCIR 2011) (Mountain View, CA, USA, 20 October 2011).)

[3] 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).

[4] Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. 2007. DBpedia: a nucleus for a Web of open data. In *LNCS 4825: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). Springer-Verlag, Berlin/Heidelberg, 722-735.

[5] Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., and Patel-Schneider, P.F. 2003. *The Description Logic Handbook: Theory, Implementation and Applications*. Cambridge University Press, West Nyack, NY.

[6] Berners-Lee, T., Hendler, J., and Lassila, O. 2001. The Semantic Web. *Scientific American* 5.

[7] Cheng, T., Yan, X., and Chang, K.C.-C. 2007. 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).

[8] Cheng, T., Yan, X., and Chang, K.C.-C. 2007. 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, 387-398.

[9] 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 *LNCS 4862: INEX 2007*, N. Fuhr et al., Eds. Springer-Verlag, Berlin/Heidelberg, 245-251.

[10] 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).

[11] Milne, D., Witten, I.H., and Nichols, D.M. 2007. A knowledge-based search engine powered by Wikipedia. In *Proceedings of the16th ACM Conference on Information and Knowledge Management (CIKM 2007)* (Lisbon, Portugal, 6–8 November 2007). ACM Press, New York, NY, 445-454.

[12] 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, 697-706.

[13] 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), 203-207.

[14] 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, 1015-1018.