Information Access & Information Extraction

Anselmo Peñas, Department of computing Languages and Systems, National Distance Learning University, Spain

TrebleCLEF Summer School on
Multilingual Information Access
Hotel Santa Croce in Fossabanda, Pisa, Italy
15-19 June 2009

Information Extraction & Information Access

Anselmo Peñas
NLP & IR Group
UNED

nlp.uned.es

What is this about?

"Information Extraction" & "Information Access"

Talk about IE
AND
Talk about IA

¿?
Talk about (IE ∩ IA)

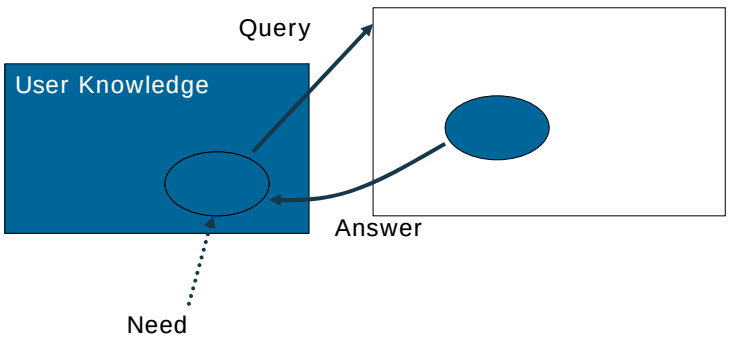
We will not talk about the semantics of the Boolean queries

Content

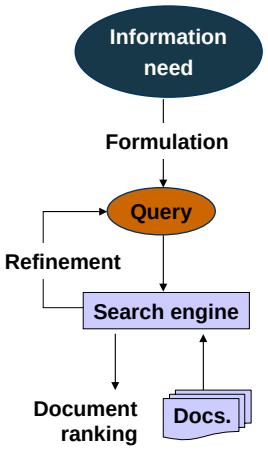
1. Information Access is something more than search
2. Search is something more than querying
3. Browsing & IE
4. Terminology Retrieval & Browsing
5. Some current evaluation tasks of IR+IE
6. Conclusion

Information Access models

First simple model: **Information need** is a hole in the knowledge we require



Classic IR model

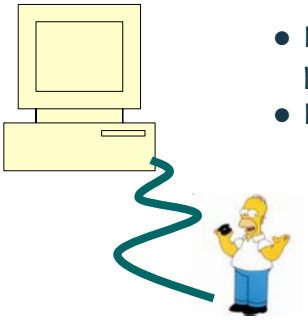


Classic IR model

- Ignores the user cognitive process
- Assumes that the information need
  - Is previously defined
  - Is static
  - Can be formulated as a query
  - The user knows how to do that
- Searching is Query Refinement
- The value is in the result, not in the process

Information need

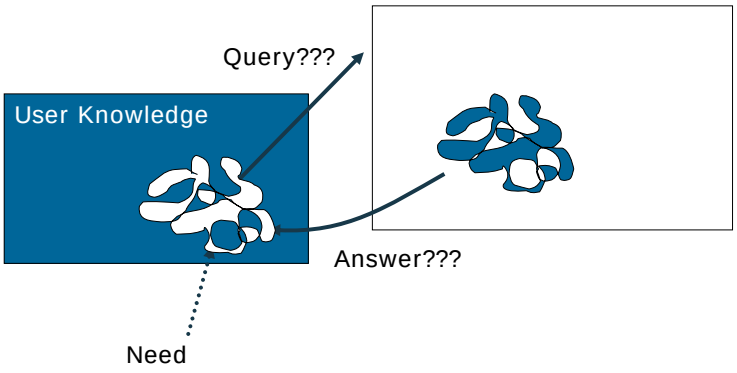
If the machine doesn't go to the human…
… the human goes to the machine



- I asked a lawyer about expressing his particular **information need**
- He answered me: "car kilometer invoice"
  - Give me something and I will manage

Information Access models

Belkin 1980: Information need cannot be accurately understood and represented



Our searching strategies

- Top-down
  - From general terms to specific key-words
  - Reducing the number of results
  - Finding the appropriate vocabulary
- Bottom-up
  - Increasing the number of results

Our searching strategies

Why do we use strategies?

Our searching strategies

Who is helping us?

Information Access models

- This is a subjective task
  - Depends on the way user assimilates information
  - Depends on the previous knowledge
  - ...
- The information need is defined and clarified along the whole process
  - We learn reading titles, finding new terms, reading snippets and documents, surfing to other pages
- It is an interactive task
- The value is in the process, not only in the result

Information Access models

- Systems should assist the whole process
- Should permit storing and managing intermediate results
- Systems should be different depending on the information need

Information Access

Nobody ask us:

What are you going to do with the information?

Some tasks after searching

- Look for Trends
- Make Comparisons
- Aggregation and Scaling
- Identify a Critical Subset
- Assess
- Interpret
- The rest:
  - cross-reference
  - summarize
  - find evocative visualizations
  - miscellaneous

O'Day & Jeffries, Orienteering in an information landscape: how information seekers get from here to there, Proceedings of InterCHI '93. Slide from M. Hearst

Information Access

Information Access is something more than searching

Information Access process

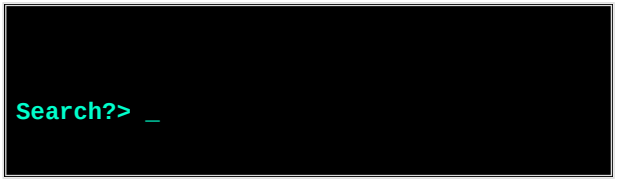
1. Determining conditions
   - Need, form and use of information
   - Barriers (language, vocabulary, ...)
2. Sources and system selection
   - Media, strategy, interaction schema
3. Locating, searching and retrieval
4. Analysis and selection
5. Synthesis
6. Evaluation
7. Revision

Content

1. Information Access is something more than search
2. Search is something more than querying
3. Browsing & IE
4. Terminology Retrieval & Browsing
5. Some current evaluation tasks of IR+IE
6. Conclusion

IR evolution (not exactly Revolution)

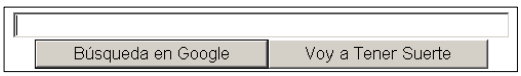
In the 50's...



Searching interfaces

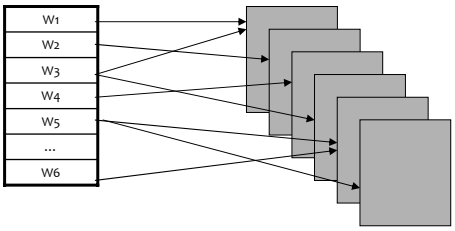
Querying

... Now



Why our interfaces tend to be like these?

Free-text indexing

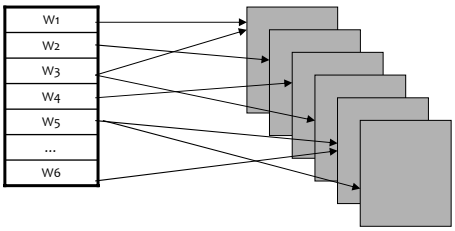


- Indexing requires
  - Entry terms to the index: QUERY
- Efficiency
  - But remember is a mean, a technical solution

Free-text indexing

Is a query the only way to access indexes?

Is an index the only way to access collections?



Free-text indexing

Searching is something more than indexing

Querying

Understood as posing some words in a text box

- Is not dynamic
- Limits the possibility of inferences
- Is a path of a single step: Query -> Result

- Limits our Information Access interfaces

Querying

- A path of a single step?
- What do search engines return?

What has changed?

```
Search?> Chrysallida Josebai_

1. Shell Catalogue
2. Family PYRAMIDELLIDAE ODOSTOMIINAE CHRYSALLIDINI
3. Chrysallida Carpenter, PP, 1856
4. Type species: Chrysallida (Parthenina) josebai

Options [1..4], Search?> _
```

What has changed?

Browsing

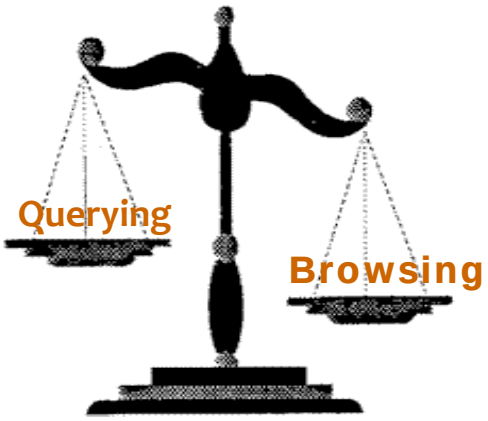
Searching

- Searching is something more than querying
- Even under the querying paradigm

Content

1. Information Access is something more than search
2. Search is something more than querying
3. Browsing & IE
4. Terminology Retrieval & Browsing
5. Some current evaluation tasks of IR+IE
6. Conclusion

More Browsing



Users

- **User optimizes effort**
- We are better recognizing than remembering
  - Menu vs. Command Line
- We prefer clicking before writing

Browsing vs. Surfing

- Clicking a link or choosing a menu option…
  Is it "Browsing?"
- No necessarily
- The key: Do we have intermediate levels of Information Access?
- Again: Does a search engine return documents?

Intermediate levels

- Which is the basic information unit of our application?
  - Final level: Document, Record, Passage, Entities, Numerical data...
- Which is the possible opening?
- Which are the intermediate levels?

Intermediate levels

- Example:
  - Information Unit: document
  - Unrestricted collection of 25.000 million documents
  - Opening?
  - Intermediate levels?
  - Think

Intermediate levels

- 2 answers
  - Opening: querying
    - Intermediate level:
      - Titles + KWIC (Keywords in Context)

Google

  - Opening: taxonomy browsing
    - Intermediate levels:
      - Taxonomy
      - Titles + document beginning / meta description

Yahoo!

Intermediate levels

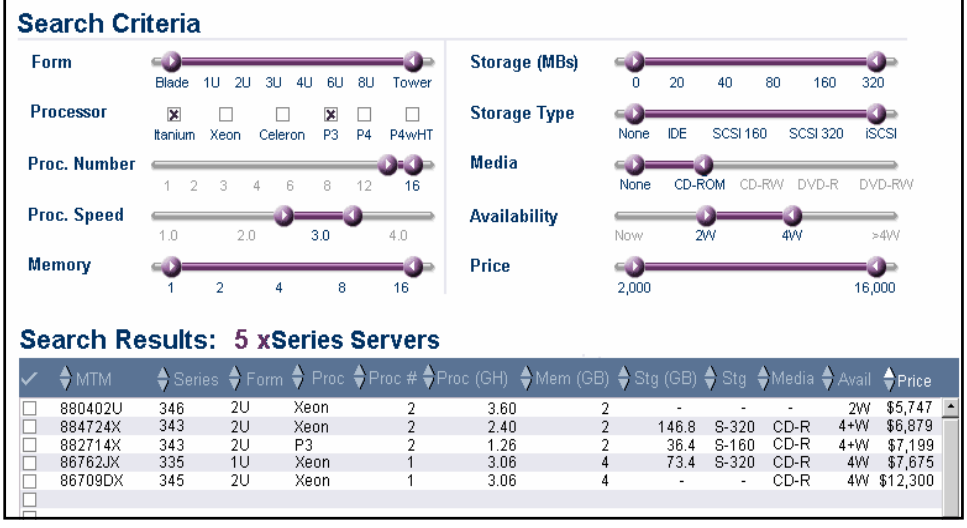
- Why not different openings?
  - **"What do you want to do"**
    - Download a resource? (Resource need)
      - Image, Music, Video...
    - Buy something? (Transactional need)
      - Product, Service...
    - Get some information or data? (Informational need)
    - Find a web site? (Navigational)
    - ...
  - **"What about?"**

Intermediate levels

More browsing, more structure

- Browsing requires structure
  - No problem with Data Bases
    - Try http://www.trovit.com

Molander et al, 2005 IBM



MTM	Series	Form	Proc	Proc #	Proc (GH)	Mem (GB)	Stg (GB)	Stg	Media	Avail	Price
880402U	346	2U	Xeon	2	3.60	2	-	-	-	2W	$5,747
884724X	343	2U	Xeon	2	2.40	2	146.8	S-320	CD-R	4+W	$6,879
882714X	343	2U	P3	2	1.26	2	36.4	S-160	CD-R	4+W	$7,199
86762JX	335	1U	Xeon	1	3.06	4	73.4	S-320	CD-R	4W	$7,675
86709DX	345	2U	Xeon	1	3.06	4	-	-	CD-R	4W	$12,300

More browsing, more structure

- Browsing requires structure
  - What about repositories?
    - Semi-structured information
    - Metadata, Facets, Tags
  - No problem if we have it
    - Automatic categorization

More browsing, more structure

- Browsing requires structure
  - What about Free-text?
    - Unstructured information
  - Create structure automatically
    - Global (static): Over the whole collection
    - Local (dynamic): Over searching results
  - Information Extraction

Information Extraction (IE)

- Entities
  - People: John Smith, J. Smith, Smith, John, Mr. Smith
  - Locations: EU, The Hague, Pisa, Piazza Nova
  - Organizations: IBM, The Beatles, University of Pisa
  - Quantities: 10 km, forty per cent, 40%, $10
- Relationships
  - [Barack Obama]$_1$ is [President$_2$ of the [United States]$_3$]$_4$
    - nation(3), president(1,3), coref(1,4)
- Events
  - Financial Events: Changes of management
  - Socio/Political Events: Traffic accidents
  - Geographical Events: Natural Disasters

Two kinds of IE approaches

Knowledge Engineering

- Rule based
- By experienced language engineers
- Make use of human intuition
- Only small amount of training data
- Development could be very time consuming
- Difficult to maintain

Learning Systems

- Developers do not need LE expertise
- Requires annotated training data (Supervised or Semi-supervised)
- Some changes may require re-annotation of the entire training corpus
- Difficult to ensure the quality of annotations

Some Hand-Coded IE Systems

- FRUMP [DeJong 82]
- CIRCUS / AutoSlog [Riloff 93]
- SRI FASTUS [Appelt, 1996]
- OSMX [Embley, 2005]
- DBLife [Doan et al, 2006]
- Avatar [Jayram et al, 2006]

Learning-Based IE

- Hidden Markov Models [Leek, 1997]
- Maximum Entropy Markov Models [McCallum et al, 2000]
- Conditional Random Fields [Lafferty et al, 2001]
- Dynamic Conditional Random Fields [McCallum, 2003]

Semi-Supervised IE Approach

Learn to Gather More Training Data

1. Use a small set of labeled data (seed) T to learn an extraction model E
2. Apply E to find mentions in document collection
3. Select some mentions to add to labeled data T'
4. Use T' to learn a new extraction model E'
5. Repeat

Example: Recipes (3500 docs)

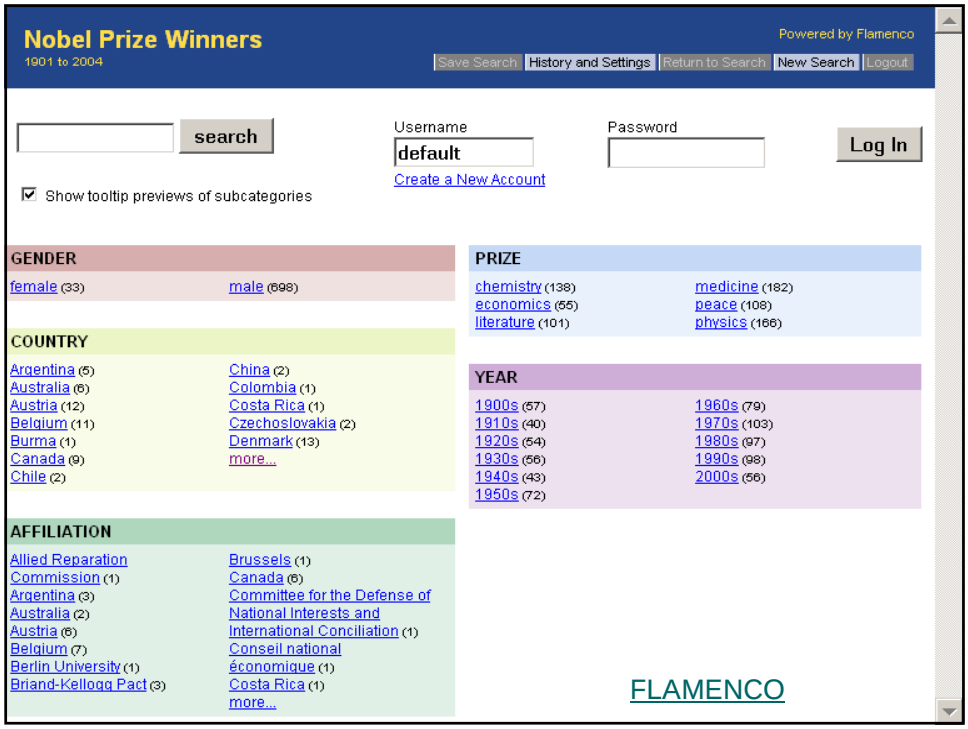
COOK-3300 ABSTRACT Molasses Brownies Recipe - Recipe for Molasses Brownies ; More Molasses Cookies ; Molasses Brownies ; 2/3 cup butter ; 2/3 cup confectioners' sugar ; 2/3 cup Grandma's Molasses ; 1 teaspoon vanilla ; 2 eggs ; 1 3/4 cup sifted all-purpose flour ; 1/8 teaspoon baking soda ; 1 cup chopped walnuts or pecans ; Cream butter and sugar until fluffy. Stir in molasses and vanilla. Beat in eggs. Sift together the flour and baking soda; add to first mixture, blending well. Stir in chopped nuts. Spread batter in two generously greased 9-inch square pans. Bake at 350° for 25 minutes. ;Makes about 50 molasses brownies.
;
COOK-3301 ABSTRACT Georgia Mint Julep Recipe - Recipe for Georgia Mint Julep with Peach Brandy ; Georgia Mint Julep ;Chill silver julep tumblers or mugs. For each tumbler, dissolve 1 teaspoon of sugar in a little water. Reserve. Fill each tumbler with finely crushed ice. Pour in 3/4 ounce brandy and 3/4 ounce peach brandy. Stir in dissolved sugar. Stir until the outside of tumbler is heavily frosted. Garnish with 5 or 6 sprigs of fresh mint, tucked in the ice so they stand out. Serve with a straw, if desired.;

From Stoica & Hearst, HLT-NAACL '07

Putting IE and Browsing together

- Example: Create Facet Hierarchies

  (Stoica & Hearst, HLT-NAACL '07)



Facets

- The case of Facets
  - Independent dimensions in the search space
  - A Facet is a range of labels organized in a hierarchy
  - An item have no more than one label per facet
- Successful idea linked to Browsing

Facets

http://search.express.ebay.com/

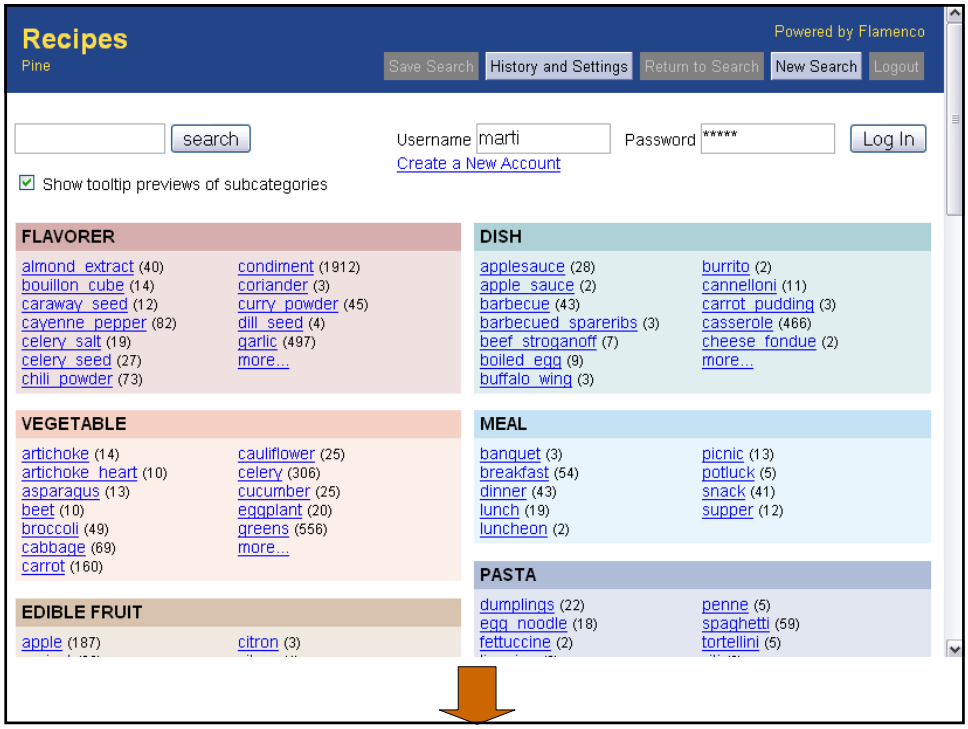
Advantages

- Give control y flexibility
  - Decide when is time for querying or for browsing
  - Choose the opening facet
  - Always under the same view
  - User can predict the behavior (not with ranking or clustering)
  - User feels that no information is lost
  - Browsing doesn't end with empty list of results
  - Easy to come back an step

Example: Recipes (3500 docs)

COOK-3300 ABSTRACT Molasses Brownies Recipe - Recipe for Molasses Brownies ; More Molasses Cookies ; Molasses Brownies ; 2/3 cup butter ; 2/3 cup confectioners' sugar ; 2/3 cup Grandma's Molasses ; 1 teaspoon vanilla ; 2 eggs ; 1 3/4 cup sifted all-purpose flour ; 1/8 teaspoon baking soda ; 1 cup chopped walnuts or pecans ; Cream butter and sugar until fluffy. Stir in molasses and vanilla. Beat in eggs. Sift together the flour and baking soda; add to first mixture, blending well. Stir in chopped nuts. Spread batter in two generously greased 9-inch square pans. Bake at 350° for 25 minutes. ;Makes about 50 molasses brownies.
;
COOK-3301 ABSTRACT Georgia Mint Julep Recipe - Recipe for Georgia Mint Julep with Peach Brandy ; Georgia Mint Julep ;Chill silver julep tumblers or mugs. For each tumbler, dissolve 1 teaspoon of sugar in a little water. Reserve. Fill each tumbler with finely crushed ice. Pour in 3/4 ounce brandy and 3/4 ounce peach brandy. Stir in dissolved sugar. Stir until the outside of tumbler is heavily frosted. Garnish with 5 or 6 sprigs of fresh mint, tucked in the ice so they stand out. Serve with a straw, if desired.;

From Stoica & Hearst, HLT-NAACL '07



Advantages

- Integrates querying and browsing
  - Querying modify the hierarchies (facets)
- Standard technology
  - Data Bases
  - Easy to add more items to the repository

Content

1. Information Access is something more than search
2. Search is something more than querying
3. Browsing & IE
4. Terminology Retrieval & Browsing
5. Some current evaluation tasks of IR+IE
6. Conclusion

Where is the money?

- Best Practices on the Use of Information Retrieval in e-Discovery
  (Sedona Conference Journal, 2007)
  TREC Legal Track
- Discovery in Law
  - Request documents and other evidence from other parties, or require the production of evidence
- eDiscovery
  - Information in electronic format

Search words?

- Best Practices on the Use of Information Retrieval in e-Discovery
  - "Simple keyword searching, while itself a valuable tool, has certain known deficiencies"
  - "Human language elasticity allows for private codes and vocabularies to exist in different subcultures in any enterprise, thus making the **identification of the *words* to be searched** much more challenging".
- A problem of recall

Language barriers

- Specific domain terminology
- Translinguality
- Terminology Variation

Terminology variation

- Everyone uses different words
  - Less than 20% of coincidence (Furnas, 1987)
  - One single good word retrieves a small portion of relevant documents
  - Content producer and content searcher are not the same person

Furnas, G.W., et al. (1987). "The Vocabulary Problem in Human-System Communication," *Communications of the ACM*, 30(11): 964-971.

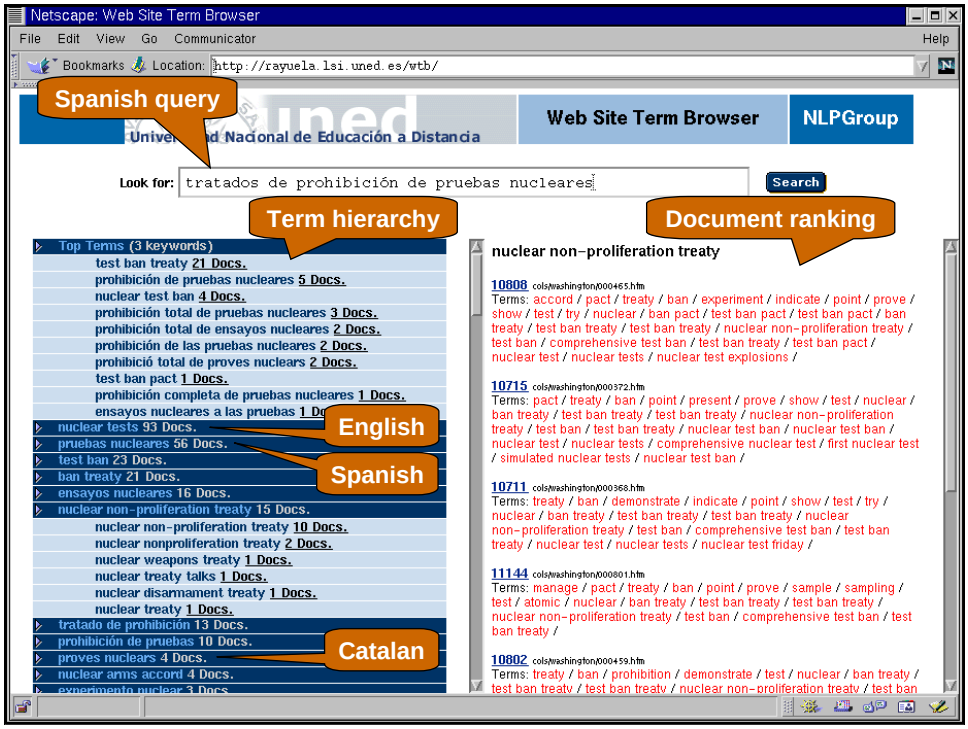
Terminology Variation

How to help user?

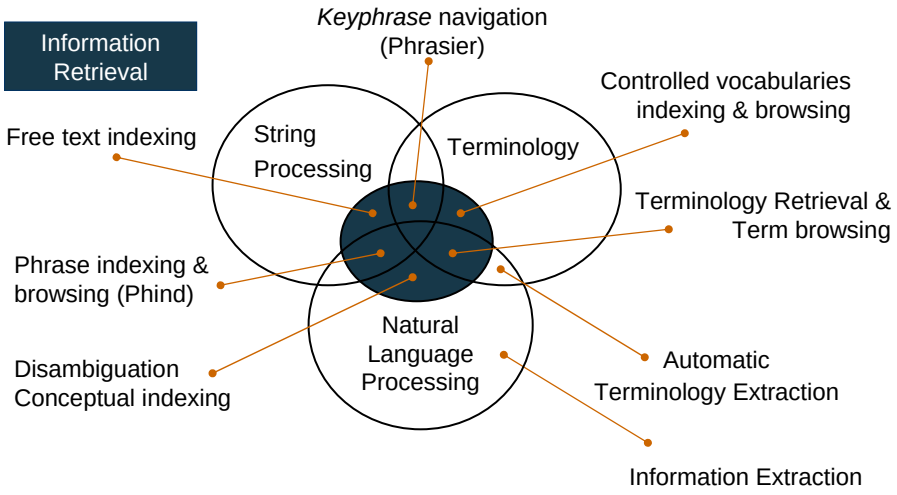
- From querying paradigm?
  - Pseudo-relevant feedback
  - Transparent for user
- From browsing paradigm?

Terminology Extraction

- Obtain terminology lists automatically
  - Language model of a generic corpus
  - Language model of a specific corpus
  - Divergence
  - Used to build thesaurus
- Terminological Phrases (lexicalized sequences of words)
  - NGRAMS
  - [N|Adj]+ [N|Adj|Det|Prep]* [N|Adj]+
  - Document Frequency
  - Almost unambiguous, easy to read, select or discard



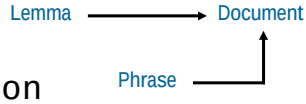
Term retrieval and browsing



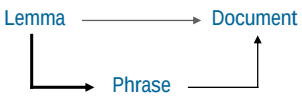
WTB: Phrase indexing

Steps

1. Text pre-processing and listing of words
2. Word tagging (oriented to phrase detection)
3. Phrase detection & lemmatization of components
4. Document indexing & statistics (document frequency)



5. Phrase selection (Subsumption & Lexicalization degree)
6. Phrase indexing



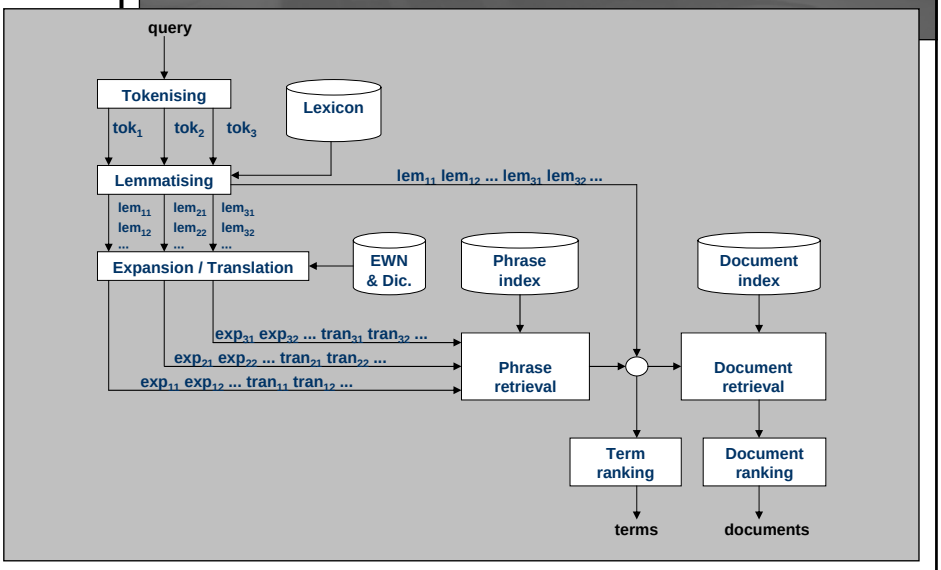
Query expansion and translation

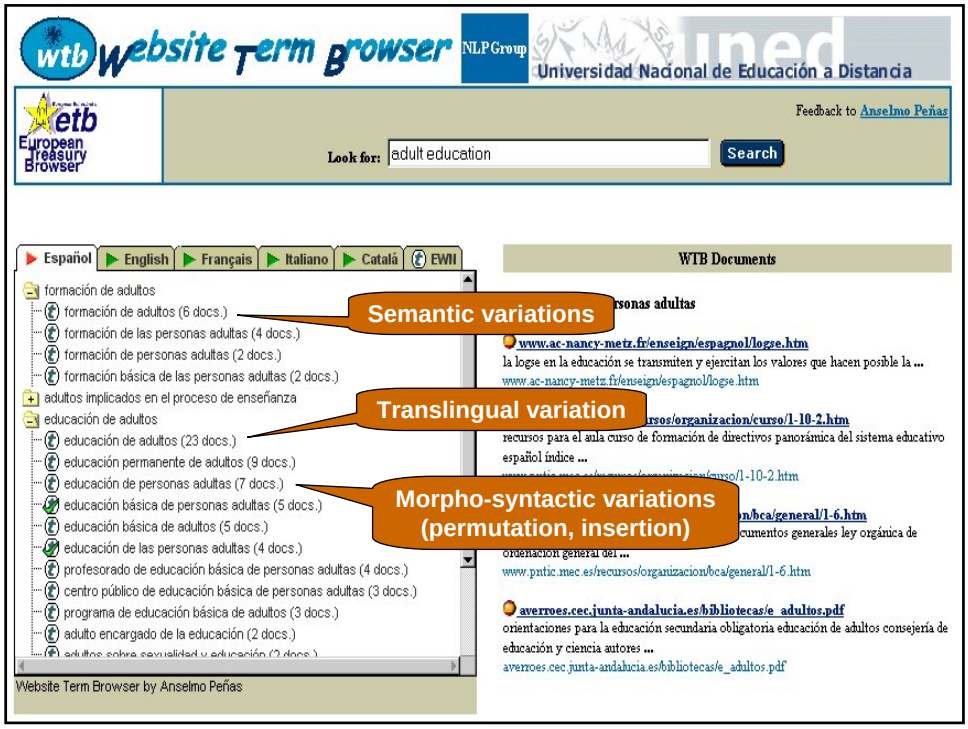
	Tratados	de	Prohibición	de	Pruebas	Nucleares
Expansion	acuerdo capitulación concertación convenio cuidar, pacto manejar procesar		embargo entredicho interdicción interdicto proscripción		cata, catadura degustación ensayo escandallo experimento gustación muestreo, tanteo	nuclear
Translation	accord discourse handle manage pact process treat treatise treaty		ban interdiction prohibition proscription		demonstrate establish, exhibit experiment experimentation fall, fitting indicate, point present, proof prove, run sample, sampling shew,show, taste test, trial, try	nuclear

Reducción ambigüedad

Nuclear taste proscription process? ✗ Nuclear test ban treaty? ✓

WTB: Retrieval





Term browsing?

Google suggest

Google Suggest LABS

Web Images Groups News Froogle Maps Scholar more »

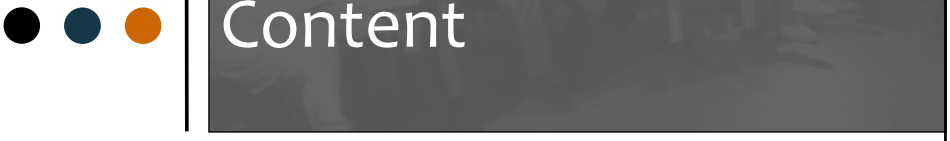
natural la

natural law	18,600,000 results
natural language processing	5,460,000 results
natural laxatives	169,000 results
natural laxative	155,000 results
natural law party	6,460,000 results
natural labor induction	204,000 results
natural language	16,700,000 results
natural law theory	5,840,000 results
natural laws	12,400,000 results
natural laxitives	2,090 results

Advanced Search
Preferences
Language Tools

As you type, G lts. Learn more

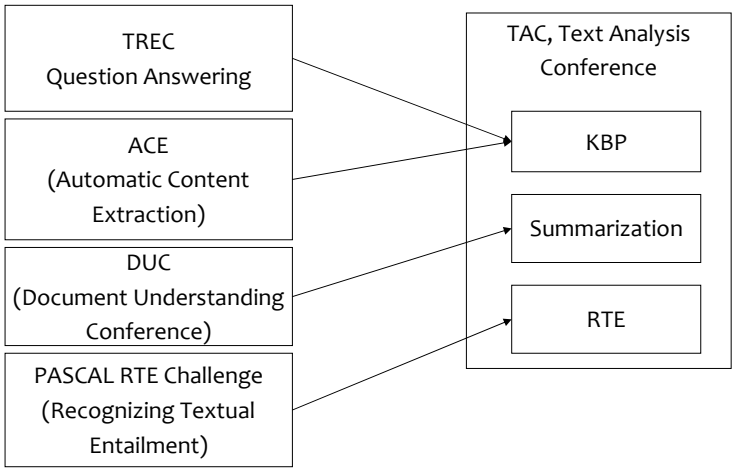
BuscaUned



Content

1. Information Access is something more than search
2. Search is something more than querying
3. Browsing & IE
4. Terminology Retrieval & Browsing
5. Some current evaluation tasks of IR+IE
6. Conclusion

Text retrieval & Text analysis



Knowledge Base Population

- Explore extraction of information about entities
- With reference to an external knowledge source
  - Basic schema for persons, organizations, and locations
  - Nodes in an ontology must be created and populated
  - Using unstructured information found in text
  - Wikipedia Infoboxes serve as a initial knowledge representation

Knowledge Base Population

- "The goal would be to link the spouse field in the 'Paul Newman' node to another ontology node -- one for Joanne Woodward -- and not merely provide a textual fragment containing her name"
- "The goal of updating an existing knowledge source will require synthesizing information from multiple documents and grounding entity mentions within the knowledge base"
- "The problem can be formulated as a QA task -- slots can be filled in by asking questions, like "Where was Paul Newman born?"

Knowledge Base Population

- Entity Linking
  - [Name_string, docid] -> KB_node_id
    - Similar to cross-document co-reference
    - Linkage to knowledge base, instead of clustering
- Slot Filling
  - Predefined set of relationships and attributes for target entities
  - [Name_string, docid, entity type, KB_node_id] ->
    ->[Slot_name, Answer String, docid_Support, target_node_id]
  - Slots
    - Single value: PER:date-of-birth
    - Multiple values: PER:employer (KB_node link)

Knowledge Base Population

Persons

- Aliases/Nicknames/Variants
- Birth name
- Age
- Birth date
- Birth place
- Home town
- Death date
- Resides-in
- Nationality(ies)
- Schools attended
- Degrees held
- Employer(s)
- Occupation
- Religion
- Spouse
- Parents
- Children
- Siblings
- Email address
- Phone number
- Salary

Knowledge Base Population

Organizations

- Aliases/Nicknames/Variants
- Date established
- Location of Headquarters
- Membership size
- Leader
- Stock ticker
- Annual revenues, budget, income
- Motto
- Website

GPEs

- Aliases/Nicknames/Variants (also previous and foreign names)
- Date settled
- Latitude and Longitude
- State or Province
- Country
- Population
- Political leader
- Seat of government (i.e., county seat or state/national capital)

Web People Search

JAVIER ARTILES (UNED)
SATOSHI SEKINE (NYU)
JULIO GONZALO (UNED)

WEPS 2 WORKSHOP, WWW 2009
MADRID, 21 APRIL, 2009

Web People Search

john tait
Google Search | I'm Feeling Lucky

What we get | **What we want**

John Tait - [Traducir esta página]
Professor **John Tait** University of Sunderland School of Computing Engineering and Technology Tom Cowie Campus at St. Peter's Sunderland ...
osiris.sund.ac.uk/~cs0jta/ - 21k - En caché - Páginas similares

John Tait (American football) - Wikipedia, the free encyclopedia - [Traducir esta página]
John Bernard **Tait** (born January 26, 1975 in Phoenix, Arizona) was an offensive lineman for the both the Kansas City Chiefs, and the Chicago Bears of the ...
en.wikipedia.org/wiki/**John_Tait**_(American_football) - 43k - En caché - Páginas similares

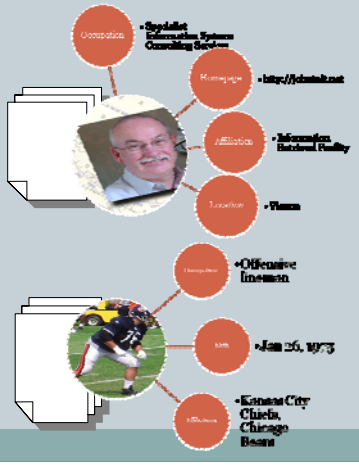
John Tait - Wikipedia, the free encyclopedia - [Traducir esta página]
John Tait (athlete), Olympic athlete; **John Tait** (American football), professional football player; **John Tait** (horseman) (1813-1888), Australian Thoroughbred ...
en.wikipedia.org/wiki/**John_Tait** - 18k - En caché - Páginas similares
Más resultados de en.wikipedia.org »

John Tait - Specialist Information Systems Consulting Services - [Traducir esta página]
Information architecture, Web Design, Web Standards.
www.**johntait**.net/ - 5k - En caché - Páginas similares

John Tait's Home Page - [Traducir esta página]
John Tait's Home Page. johntait.net specialist consultancy ... very proud of the PhD students I have supervised, who include: **John** Macintyre Siobhan Devlin ...
www.**johntait**.net/home.html - 10k - En caché - Páginas similares

John Tait Stats, News, Photos - Chicago Bears - ESPN - [Traducir esta página]
Bears March 12: Signed cornerback Marcus Hamilton to a one-year contract; placed tackle **John Tait** on the reserve/retired list. ...
sports.espn.go.com/nfl/players/profile?playerId=1765 - 47k - En caché - Páginas similares

John Tait Numeritos, Noticias, Fotos - Chicago Bears - ESPN Deportes



Is this really what we want?

Users can refine their query

- "John Tait IRF" → IR researcher homepage

Promoting diversity in search results is a broader solution

- First hits should be about different John Tait

Yes, it is often what we want

Web person profiling

80% U.S. companies check the web before hiring someone

- In 30% cases web results impact hiring decision (source: notoriety.com).

Most Web users do it! (at least egosurfing)

- (raise your hand if you haven't ☺)

Popularity & reputation management

What we get

Undetected coreference

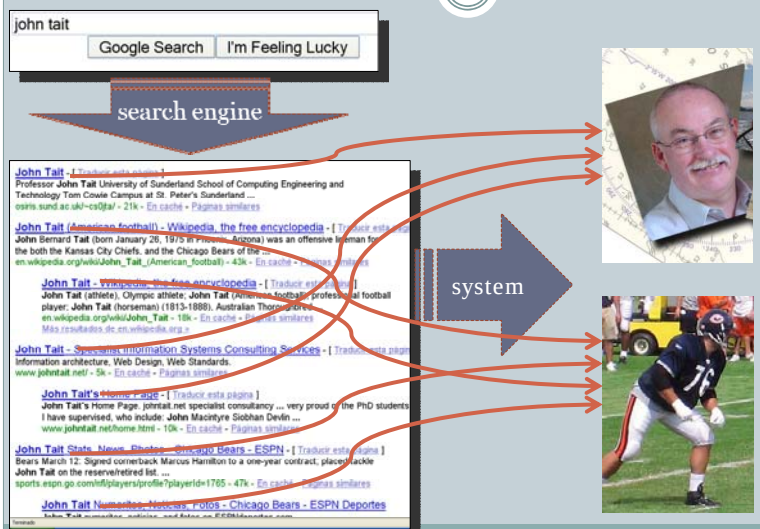
Unresolved ambiguity

How relevant is this problem?

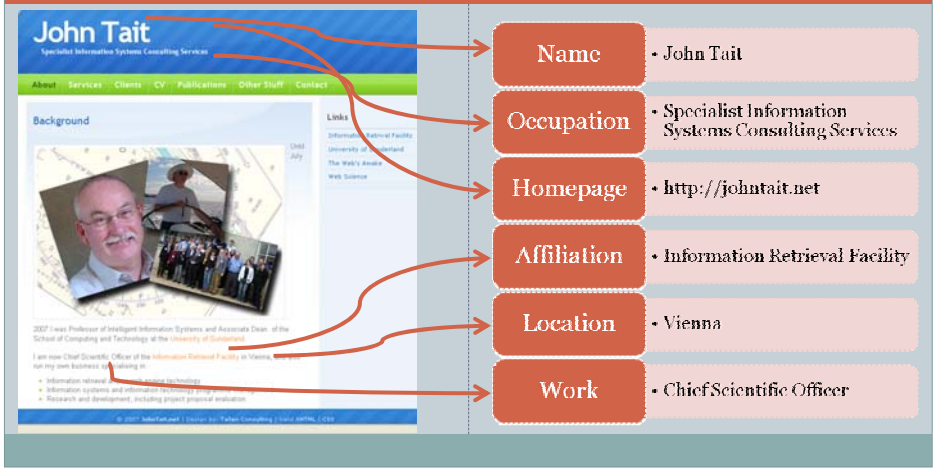
- 11-17% of Web queries include a person name
- 4% of Web queries are just a person name
- U.S. Census Bureau: 90,000 names shared by 100,000,000 people
- Wide variations in ambiguity for search results
- Web People Search engines available since 2005

2. WePS 2 Tasks

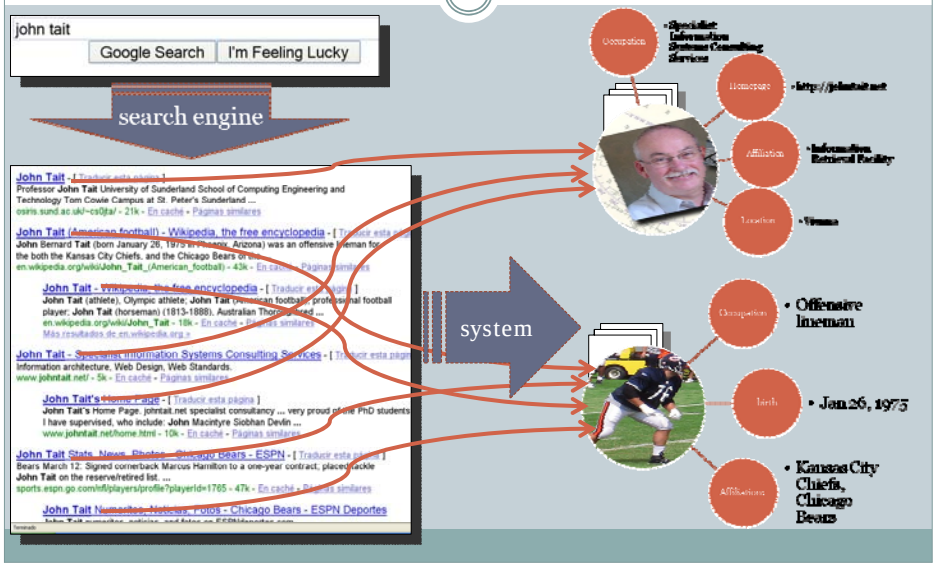
WePS clustering task



WePS Attribute Extraction Task



Full WePS task (not evaluated!)



Attributes

Occupation, affiliation & work are the most common

Most attributes appear in less than 1/10 of the documents

	Attribute Class	Examples of Attribute Value	Total Number	Average per doc	Max. per doc
1	Date of birth	4 February 1888	370	0.12	4
2	Birth place	Brookline, Massachusetts	301	0.10	4
3	Other name	JFK	797	0.27	6
4	Occupation	Politician	3,292	1.10	20
5	Affiliation	University of California, Los Angels	3,105	1.03	19
6	Work	The Secrets of Doroon	3,770	1.25	141
7	Award	Pulitzer Prize	264	0.09	14
8	School	Stanford University	494	0.16	10
9	Major	Mathematics	173	0.06	6
10	Degree	Ph.D.	335	0.11	6
11	Mentor	Tony Visconti	343	0.11	12
12	Location	London	888	0.30	8
13	Nationality	American	250	0.08	2
14	Relatives	Jacqueline Bouvier	914	0.30	29
15	Phone	+1 (111) 111-1111	219	0.07	5
16	FAX	(111) 111-1111	65	0.02	2
17	Email	xxx@yyy.com	209	0.07	5
18	Web site	http://nlp.cs.nyu.edu	154	0.05	4

Attribute Extraction Results

Difficult task!

Table 5. Results by System

System	Precision	Recall	F-measure
PolyUHK	30.4	7.6	12.2
ECNU_1	6.8	18.8	10.0
ECNU_2	8.0	17.6	11.0
MIVTU	5.7	15.5	8.3
CASIANED	8.5	19.0	11.7
UC3M_1	2.5	2.2	2.3
UC3M_2	2.4	2.2	2.3
UC3M_3	2.2	2.0	2.1
UC3M_4*	2.2	2.0	2.1
UC3M_5*	8.0	3.6	5.0
UvA_1	2.7	27.3	5.0
UvA_2	4.4	27.4	7.6
UvA_3	0.7	0.2	0.2
UvA_5	0.2	0.0	0.0
UvA_5	3.3	2.8	3.1

* indicates unofficial runs, which are sent after the deadline

Different Attributes, Different Results

4 types of attributes based on the characteristics

Attribute	Description	Performance	Comments
Phone, FAX, email, Website	There is a typical pattern	R: 74-40 (ECNU, UvA)	Disambiguation is needed.
Degree, Nationality	Unfamiliar NE, but candidates are limited	R: 43-42 (CASIANED)	We need a good NE tagger for the category. Maybe possible.
D. Of birth, Birth place, Other name, Affiliation, School, Mentor, Relative	Typical NE, disambiguation is needed	R: 55-17 (MIVTU, UvA, PolyUHK)	NE tagger is ready. We need good disambiguation
Award, Major, Occupation	Unfamiliar and difficult NE type	R: 17-38 (UvA)	We need a good NE tagger for the category. It looks very difficult

Typical System Strategy

Most systems use two phase strategy

1. Find the candidates
   - Use NE tagger, gazetteer, regular expression to find candidates which have the same type to the target attribute
2. Filter (verify) the candidates
   - Select only those which are the attribute-values of the target person. It can be done by local pattern, supervised classification, distance & cue phrase.

Conclusion

1. Information Access is something more than search
2. Search is something more than querying
3. Browsing & IE
4. Terminology Retrieval & Browsing
5. Some current evaluation tasks of IR+IE