Semantic annotation and expansion for keyword queries part 1, tutorial

Ranka Stanković
ranka@rgf.rs
University of Belgrade, Serbia

http://rgf.rs/ranka1.pptx
University of Belgrade is the oldest State university of the Republic of Serbia.

The origin of the University of Belgrade can be tracked down to the beginning of the 19th century, when Dositej Obradović founded the College in 1808.

http://bg.ac.rs/en

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, Univesiy of Santiago de Compostela, Spain, 18-22 July 2016
University of Belgrade, Serbia

- 31 Faculties
- 11 scientific research institutes
- 6 University Centers
- University Library
- 90,000 students and 2,650 teachers

- More than 7000 staff;
- More than 320 modules;
- More than 1700 students enrolled at master and PhD programs;
- More than 350,000 graduated students, 23,000 MSc and 13,500 PhD fellows
Society was established in order to achieve the objectives in the field of promotion, popularization of all branches of linguistic technology in scientific, professional and practical level.

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
Main topics

• Introductory definitions
  ▫ Information need & Information Access
  ▫ Query & Semantic annotation

• A review of ontology based query expansion
• A keyword-based semantic retrieval approach
• Interaction Between Automatic Annotation and Query Expansion
• Query by (lexical) pattern
• Examples of application on parallel and domain specific (math) corpora
Motivation

- Enhancing the search results in large archives is a concern shared by many research community
- The improvement can come from two directions:
  - enhancing the annotations or
  - enhancing the search mechanism.
- Both directions are active research area’s.

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, Universiy of Santiago de Compostela, Spain, 18-22 July 2016
Information Access

Definition [Wikipedia]

• Area of research at the intersection of Informatics, Information Science, Information Security, Language Technology, Computer Science, and Library Science.
• The objective [...] is to simplify and make it more effective for human users to access and further process large and unwieldy amounts of data and information.

Technologies

• Information Retrieval, Text Mining,
• Machine Translation, and Text Categorisation.

Related terms

• search, information retrieval, searching,
• navigation, data exploration, quering, ...
Information Access

Existing approaches

Query Languages (QL)
- searching, querying

Navigation Structures (NS)
- navigating, browsing, exploring

Interactive Views (IV)
- interacting, dialoging, selecting, composing, transforming

Ranka Stanković. "Semantic annotation and expansion for keyword queries". 2nd KEYSTONE Training School on Keyword search in Big Linked Data, Santiago de Compostela. 18-22 July 2016
Query Languages

- The user provides some input query, and the system returns some answers to the query

- Information retrieval (IR) search: keywords, forms
- Corpus query languages: CQP, CQL, regular expressions
- Formal query languages: SQL, XQuery, SPARQL
- Natural language interfaces (NLI): IBM Watson
Query Languages

1. Which type of questions can be answered?
2. Whether users are guided in the expression of their information needs?
3. How users comprehend user interface components and controls?
4. What amount of data can be accessed with acceptable response times?

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, Univesiy of Santiago de Compostela, Spain, 18-22 July 2016
Information need

- An incomplete query disrupt a search engine from satisfying the user’s information need.
- The query provided by the user is often unstructured and incomplete.
- We need some representation which can express the user’s information need.
What is Query Expansion?

**Query Expansion**
- adding search terms to a user’s keyword list

**The goal is**
- to improve precision and/or recall.

**Simple example**
- User Query: “car”
- Expanded Query: “car cars automobile automobiles auto” etc...

Ranka Stanković. "Semantic annotation and expansion for keyword queries”. The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
Why Query Expansion?

- Query expansion is very important on the web.
- The amount of information on the web is always increasing.
- Search engine users follow specific trends with their searches.
  - 2-3 words
  - Broad search term
  - Do not like to expand their queries either through refining search terms or using Boolean operators

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSSTONE Training School on Keyword search in Big Linked Data, Universiy of Santiago de Compostela, Spain, 18-22 July 2016
Techniques of query expansion

- Finding synonyms of words, and searching for the synonyms
- Finding various morphological forms of words by stemming or inflecting words in the search query
- Fixing spelling errors and automatically searching for the corrected form or suggesting it in the results
- Re-weighting the terms in the original query
- Finding translations of query terms (for multilingual search)
- Creating a dictionary of expansion terms for each term, and then looking up in the dictionary for expansion

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
Query expansion techniques

Ashish Kankaria, Query Expansion techniques
Indian Institute of Technology Bombay, Mumbai
The query is expanded using some external resource like WordNet, lexical dictionaries or thesaurus. Techniques involve looking up in such resources and adding the related terms to query. These resources are built manually, they contain mappings of the terms to their relevant-related terms.

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
Thesaurus based expansion

A thesaurus

- is a data structure that lists words grouped together according to similarity of meaning (containing synonyms and sometimes antonyms), in contrast to a dictionary, which provides definitions for words
- is used to expand the query terms and all the connected words of query terms are added to query.

Thesaurus based system

- have been explored and put to use by many organizations.
- example is Unified Medical Language System (UMLS) used with MedLine for querying the bio medical research literature.
- Controlled vocabulary contains similar terms for each bio medical concept.

A thesaurus based query expansion system works well only for rich domain specific thesaurus.

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, Universiy of Santiago de Compostela, Spain, 18-22 July 2016
Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, Universiy of Santiago de Compostela, Spain, 18-22 July 2016

http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus
WordNet based expansion

- WordNet is a lexical database for multiple languages.
- The similar terms from multiple languages are connected via synsets (set of senses).
- WordNet can be used to fetch related term for a particular term in multiple languages and can help in satisfying user’s information need.
- Expansion can use synonyms and/or hypernyms and/or hypernyms.

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTOKE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
Global WordNet

- A free, public and non-commercial organization that provides a platform for discussing, sharing and connecting wordnets for all languages in the world.
  - Organizes GWA Conferences – 8 conferences up to now
- Global WordNet Grid - which is being build around a shared set of concepts used in many wordnet projects.
  - List of all wordnets in the world (contact persons, licences etc. http://globalwordnet.org/wordnets-in-the-world/)
Concepts recognized by all Balkan languages

<table>
<thead>
<tr>
<th>Language</th>
<th>καδαιφ</th>
<th>халва</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bulgarian</td>
<td>κανταϊφι</td>
<td>χαλβας</td>
</tr>
<tr>
<td>Greek</td>
<td>cataif</td>
<td>halva</td>
</tr>
<tr>
<td>Romanian</td>
<td>καδαιφ</td>
<td>алва</td>
</tr>
<tr>
<td>Serbian</td>
<td>kadayif</td>
<td>kağıt helva</td>
</tr>
</tbody>
</table>

Usage of wordnets

- Improve recall of textual based analysis:
  - Query → Index
    - Synonyms: commence → begin
    - Hypernyms: taxi → car
    - Hyponyms: car → taxi
    - Meronyms: trunk → elephant
    - Lexical entailments: used a gun → shot
  - Inferencing:
    - what things can be used for transport?
  - Expressions in language generation and translation:
    - alternative words and paraphrases
Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016

http://sm.jerteh.rs

Serbian semantical resources
WordNet XML representation
Using WordNet Knowledge for Irony Classification

- A language dependent model for classification of statements into ironic and non-ironic.
  - language resources: morphological dictionaries, sentiment lexicon, lexicon of markers and a WordNet based ontology.
  - features: antonymous pairs obtained using the reasoning rules over the SrpWN (R), antonymous pairs in which one member has positive sentiment polarity (PPR), polarity of positive sentiment words (PSP), ordered sequence of sentiment tags (OSA), POS tags and irony markers (M).
- Evaluation on a collection of tweets that had been manually annotated according to irony.
- The collection of tweets is in the Serbian language (or Bosnian/Croatian/Montenegrin).
- The best achieved accuracy of the developed classifier $\text{acc} = 86.1\%$ was achieved with the set of 5 features — (PPR, PSP, POS, OSA, M).
Architecture of the ironic tweets classifier

HrTal2016, accepted, Dubrovnik September 2016
Mladenovic M., Krstev C., Mitrovic. J. Stankovic R.
„Using WordNet Knowledge for Irony Classification“
Bilingual digital library search demo in hands-on session

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
Query logs based expansion

- Query logs are maintained by search engine in order to analyze the behavior of the user while interacting with search engine.
- Query logs can be used to analyze the user’s preference and adds corresponding terms to query.
- Method fails when user search something which is not related to earlier searches.
- List of all the documents visited for a particular query can be stored for further use.
- It can be used to learn associations by combining evidence from various lexical sources like WordNet.
Relevance feedback based expansion

**Process**
- execute the initial query on collection and extract top k documents.
- use ranked document to improve the performance of retrieval.

**Assumption**
- initial retrieved documents are relevant and thus can be used to extract expansion terms.
- the initial retrieval algorithm of search engine is good.

**Model types**
- Explicit feedback from user
- Implicit feedback
- Pseudo Relevance Feedback (PRF)
Explicit feedback from user

Interactive approach: initial retrieved documents are presented to user and the user is asked to select the relevant documents.

These models are not much useful because users expect the system to be autonomous and retrieve the results for the user.

User would ultimately get irritated by repeated interaction required from him for each search.

These type of models can be used for testing search engines where developers are willing to interact with the system.


Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
Implicit feedback

• User’s feedback is inferred by the system.
• The feedback can be inferred from user’s behavior like:
  ▫ The pages which user opens for reading, or
  ▫ pages on which user clicks once the results are displayed back to the user
  ▫ Time spent on page
Pseudo Relevance Feedback (PRF)

**Process**
- Initial query is fired and top k results are obtained.
- Important terms, mostly based on co-occurrence, from these documents are extracted and added to query.
- Expanded query is re-fired to retrieve final set of documents which are made available to the user.

**Note**
- The relevancy of expansion terms depends upon the initial retrieved documents.
- Pseudo relevance feedback captures the important terms only based on co-occurrence (not enough for correctness of results)
- Semantic and lexical properties of word should be considered.
- Feedback is independent of the user there is a chance of topic drift.

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, Universiy of Santiago de Compostela, Spain, 18-22 July 2016
Pseudo Relevance Feedback (PRF)

- Typical work flow for PRF (Blind Feedback) based information retrieval system

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
Pseudo Relevance Feedback (PRF)

- Automates the manual part of relevance feedback and has the advantage that assessors are not required.
- Successfully applied in various IR frameworks and has been proved to improve precision and recall of search engines.
- Improvement of PRF by -
  - refining relevant document set
  - refining the expansion terms from PFR
  - using selective query expansion
  - varying the importance of documents

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
PFR improvements

- terms in the document which are nearer to the query terms are assigned more weight.
- use of an assisting language (AL) to improve the performance retrieval of search engine.
  - translate the query to an assisting language
  - perform PFR twice, once for in QL and other in AL
  - Merge the expansion terms obtained from both the PRF instances using translation and retrieve the documents for expanded query.
  - multilingual PRF using English as AL for French, German, Hungarian,...

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
PFR improvements

- Takes into account the structure of the documents while assigning priorities to the expansion terms.
- The intuition behind the idea is that a term that occurs in title section of a document is more important for that document than the term which occurs in the body.
- Title more compactly represents the entire document and thus it is very less probable that title will have a word which is unrelated to the document.
- Examples:
  - Project collection: Institution, Project Name, Location, Domain, Responsible person

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, Universiy of Santiago de Compostela, Spain, 18-22 July 2016
Query expansion using Wikipedia

• Wikipedia query expansion is based on the category assignments of its articles.
  ▫ The base query is run against a Wikipedia collection and each category is assigned a weight proportional to the number of top-ranked articles assigned to it.
  ▫ Articles are then re-ranked based on the sum of the weights of the categories to which each belongs.

• Thesaurus can be produced from Wikipedia articles (with some irrelevant results).

• Category information
  ▫ can be used by calculating distances between document categories and target categories.
  ▫ has more value than link information.
Query Expansion Issues

- Two major issues
  - Which terms to include?
  - Which terms to weight more?
- Concept-Based vs. Term-Based Query Expansion
  - Is it better to expand based upon the individual terms in the query, or the overall concept of the query?
- Classes of QE
  - Manual approach - Human generated thesauri
  - Interactive Query Expansion
  - Automatic Query Expansion
Approaches to Query Expansion

- Global Analysis
  - Considers all the documents in the system.
- Local analysis
  - Uses some initially retrieved documents for expansion terms.

Another classification:
- Document-term based approach.
- Query-term based approach.
- Combined approach.
Global Analysis

- Term clustering
- Latent Semantic Indexing
- Similarity Thesauri

Disadvantages
- Corpus wide statistical analysis takes computation time.
- Cannot address term mismatch problem.
The Need For Thesauri

• Naturally assumed that pulling words from a thesauri would increase:
  ▫ The number of documents retrieved.
  ▫ Possibly precision.
• The car example: “car” vs. “car, auto, automobile, vehicle, sedan, etc…”
  ▫ Which would retrieve the largest number of documents?
  ▫ Is larger necessarily better?
Human and Automatically Generated Thesauri

Earliest work began in the 1950s.

- H.P. Luhn
- *Thesaurofacet* – detailed list of engineering terms

Largely used in

- Medicine,
- Agronomy,
- Natural science,
- Technological fields.
Drawbacks of Handcrafted Thesauri

- **Cost**
  - Development.
  - Maintenance.
  - Cost often outweighs benefit.
- **Time**
  - It often takes a long time for thesauri to develop.
  - Hard to keep up with the pace of scientific and technological development.
Automatically Generated Thesauri

• Global analysis method.

• 3 Steps.
  ▫ Extract word co-occurrences or syntactic patterns.
  ▫ Define word similarities.
    • Based upon word co-occurrence or lexical relationship.
  ▫ Cluster words based upon their similarities.

• Not proven very successful.
  ▫ As late as 1990 many industries were still using handcrafted thesauri.
Interactive Query Expansion

- Uses a thesaurus.
  - After initial query is submitted,
  - the system returns a list of associated and relevant
  - words derived from both the result set and a
    thesaurus.
- Useful, but more research is needed.
Relevance Feedback

- Local analysis + interactive.
- Significant improvement in recall and precision over early query expansion work.
- Basic process as follows.
  - The user creates their initial query which returns an initial result set.
  - The user then selects a list of documents that are relevant to their search.
  - The system then re-weights and/or expands the query based upon the terms in the documents.
Automatic Query Expansion

The process of automatic query expansion using computer generated thesauri.

Works somewhat like pseudo-relevance feedback.
**Pseudo-relevance Feedback**

- Also known as blind feedback.
- Grew from problems involved in implementing relevance feedback systems.
- Users do not like to give manual feedback to the system.
Pseudo-relevance Feedback Process

1. The system returns an initial set of documents.
2. The system assumes that the top $n$ number of documents are relevant to the query.
3. The system takes terms from these documents to re-weight the query.
4. Relies largely on the system's ability to initially retrieve relevant documents.
5. May lead to “query drift.”
Concept Based Query Expansion

- Uses terms that are closer to the concept of query rather than individual query terms.
- Determining concept representing a query is hard.
- Mathematical approach

Mining for Query Expansion

- Needs a log of queries fired and the corresponding documents clicked by the user.
- If a set of documents is often selected for the same queries, then the terms in this document are strongly related to terms in the queries.
- Takes advantage of user judgment implied in the logs.
- Described in the paper

Cui, H.; Wen, J.R.; Nie, J.Y; and Ma, W.Y. 2003. Query Expansion by Mining User Logs. IEEE Transactions on Knowledge and Data Engineering.
Unitex

- http://www.igm.univ-mlv.fr/~unitex

Unitex/GramLab is an open source, cross-platform, multilingual, lexicon- and grammar-based corpus processing suite

Unitex/GramLab 3.1 Stable is now available

Unitex/GramLab has been selected as a Google Summer of Code 2016 mentor organization

Google Summer of Code (GSoC) is a global program that offers students stipends to write code for open source projects during summer break. This year, Unitex/GramLab has been selected as a Google Summer of Code mentor organization. If you're interested in helping with GSoC, mentoring a student, or you are a student, we'd love to hear from you:

- Our organisation profile
- View our ideas list
- Google Summer of Code 2016 website
- More organisations in the "languages" category
- GSoC how it works
- GSoC timeline
- GSoC FAQ

If you have any questions, please do not hesitate to post back at the users forum or to send a message to the developers mailing list.

On the Unitex/GramLab forum, you can ask and answer questions and post your suggestions about Unitex and GramLab.

Unitex/GramLab Forum

According to a study based on 377 job offers for NLP engineers from March 2013 to July 2015, Unitex is among the most expected skills in terms of NLP tools (9%)
Query in Unitex

- Regular expressions
- `<to be>` - all word forms linked to lemma
- `<N>` - search for POS – all nouns
- `<DET><A><N+Hum>` - noun preceeded by determiner and adjective
- `<N+Hum>` - nouns with semantic tag Human
- `<N+NProp+Hum~Inh>` - ...not Inhabitans
- `(<A>+<PRO+ProA>) <love>`
- `(><this>+<that>) <A> <N>`
Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
About CQPweb

The IMS Open Corpus Workbench (CWB)

The IMS Open Corpus Workbench (CWB) is a collection of open-source tools for managing and querying large text corpora (ranging from 10 million to 2 billion words) with linguistic annotations. Its central component is the flexible and efficient query processor CQP.

The first official open-source release of the Corpus Workbench (Version 3.0) is now available from this website. While many pages are still under construction, you can download release versions of the CWB, associated software and sample corpora. You will also find some documentation and other information in the different sections of this site.

The scheduled release date for CWB v3.0 was April 1st, 2010. Since then, we have moved on to add the two most-demanded features (Windows compatibility and Unicode support) in versions 3.1 and 3.2, and we are working towards a new stable release version 3.5. We welcome all interest in beta-testing these newer versions.

http://cwb.sourceforge.net/cqpweb.php

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, Universiy of Santiago de Compostela, Spain, 18-22 July 2016
About CQPweb

- **CQPweb** is a web-based app for **CQP** query processor.
- CQPweb is designed to replicate the user-interface of BNCweb tool, which also uses CQP as a back-end.
- Unlike BNCweb, CQPweb can be used with *any* corpus.
- CQPweb is especially suitable for students, non-linguists
- CQPweb can be used in three ways.
  - Via a public server. There are many of these out there; the one run by Andrew Hardie, CQPweb's main developer, is [https://cqpweb.lancs.ac.uk](https://cqpweb.lancs.ac.uk).
  - By getting a copy of the code and installing it directly on your own computer
  - By downloading CQPwebInABox, a Virtual PC which has CQPweb pre-installed (with two sample corpora included!)
Terminology extraction

- Terminology mining, term extraction, term recognition, or glossary extraction, is a subtask of information extraction.
- The goal of terminology extraction is to automatically extract relevant terms from a given corpus.
  - Used in topic-driven web crawlers, web services, recommender systems, etc.
  - Essential to the language industry
  - Used for conceptualizing a knowledge domain or for supporting the creation of a domain ontology or a terminology base
  - Used for semantic similarity, knowledge management, human translation and machine translation, etc.
- One of the first steps to model the knowledge domain is to collect a vocabulary of domain-relevant terms, as linguistic view of domain concepts.
- Automatic term extraction includes
  - Linguistic processing (part of speech tagging) to extract candidates, i.e. noun phrases, NPs (e.g. credit card", adjective-NPs "local tourist information office", and prepositional-NPs "board of directors").
  - Filtering the candidate list using statistical and machine learning methods.


Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
Examples of term extraction

- **AXN** – an adjective followed by a noun; the adjective and the noun have to agree in all four grammatical categories; e.g. zemni gas ‘natural gas’.
- **AXAXN** – a noun preceded by two adjectives that agree with it in gender, number, case and animate-ness; e.g. površinski istražni radovi ‘surface exploration works’.
- **NNgiPrepNp** - a noun followed by a noun in the genitive case and a prepositional phrase (as in case 4b); e.g. priprema ležišta za otkopavanje ‘deposit preparation for mining’.
- **NNgiNgiNgi** - a noun followed by three nouns/adjectives in the genitive case; e.g. istraživanje ležišta mineralnih sirovina ‘exploration of mineral deposits’.
- **NprepNpNgi** - a noun followed by a prepositional phrase; e.g. bakar sa primesama zlata ‘copper with a sprinkling of gold’.
- **2XAXN** - an adjective followed by a noun that agrees in all four grammatical categories and preceded by a word that does not inflect in the MWU; e.g. magmatsko-eruptivni masiv ‘magmatic-igneous massif’.
FST for extraction of MWUs of type AXN

- Two paths from one of the subgraphs that illustrate the agreement between adjectives and nouns
- Dictionary variable used for FST output in the form $a.LEMMA$ retrieves a lemma of recognized word form $a$ thus performing the simple word lemmatization
Interaction Between Automatic Annotation and Query Expansion

Search system can be implemented for collection of:
- Textual documents (monolingual, bilingual, multilingual)
- Multimodal documents (image, video)
- Georeferenced resources (point, line, polygons)

Improving a search system in two ways:
- by enriching the annotations, or
- by enriching the query mechanism

Both operations possibly benefit from preliminary
- terminological and
- lexical enrichment

Veronique Malais, Laura Hollink, and Luit Gazendam, The Interaction Between Automatic Annotation and Query Expansion: a retrieval experiment on a large cultural heritage archive, SemSearch 2008, CEUR Workshop Proceedings, ISSN 1613-0073, online at CEUR-WS.org/Vol-334/

Ranka Stanković, "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, Universiy of Santiago de Compostela, Spain, 18-22 July 2016
Enriching the annotations

- Manual annotation leads to a low number of keywords per document and improvement can be in:
  - Facilitating manual creation of annotations
  - Creation of semi-automatic annotations
  - Automatically created annotations.
- Tools can be used for semi-automatic semantic annotation, extracted from text resources.
- Automatically generated annotations seldom reach the quality level of manual annotations.

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
Annotation and Query expansion

- Retrieval of not only documents that match the query concept, but also documents that are annotated with concepts that are related to the query.
- Ontology based query expansion
- What is the effect of query expansion in the context of automatic annotation?
- Is query expansion still beneficial when applied to lower-quality automatic annotations?
- And is it still necessary if a larger number of annotations is generated?
- Case study:
  1. Compute a baseline by querying a corpus of hand-made metadata.
  2. Query the automatically generated annotations of the same corpus.
  3. Query the hand-made metadata using query expansion.
  4. Query the automatically generated annotations using query expansion.

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, Univesiy of Santiago de Compostela, Spain, 18-22 July 2016
Creating annotations

- The term “annotation” implies, very generally speaking, to attach data to some other piece of data.
- Applied to different domains
  - Document annotations
  - Semantic Wikis
  - Semantic Blogs
  - Tagging
- Annotations create a relationship between URIs and build up a network of data.
- The Semantic Web is about shared terminology, achieved through consistent use of URIs.

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016

Document annotations

Annotations can be
- manual (performed by one or more people),
- semi-automatic (based on automatic suggestions),
- or fully automatic.

Manual annotation tools
- allow users to add annotations to web pages or other resources, and share these with others.
- An example annotation would relate the text “Santiago de Compostela” to an ontology, identifying it as a city, as a capital of Galicia.

Automatic tools
- can perform similar annotations (such as named-entity recognition) without manual intervention.

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
Semantic Wikis

(Regular) Wikis
- Are collaborative hypertext authoring environments for collaborative writing and editing
- enable users to describe resources in natural language,

Semantic Wikis
- allow users to make formal descriptions of resources by annotating the pages that represent those resources.
- enable users to additionally describe resources in a formal language.

Adding metadata
- to ordinary Wiki content improves
- retrieval, information exchange, and knowledge reuse

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, Univesiy of Santiago de Compostela, Spain, 18-22 July 2016
Semantic Blogs

Blogs (or weblogs)

- are online journals or diaries, usually individual posts, created and presented in reverse chronological order.

An annotation in blogs

- a statement about a post or a category
- For example: classify posts with categories like “sports”, “cinema” or “Novak Djokovic”

Semantic Blogging

- annotations are extended, and allow association on an ontological basis.

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
Tagging

- Tagging systems (e.g., Flickr, Facebook, Linkedin,...) allow users to add tags to web resources.
- Tags express some unspecified relation between the resource and whatever the term refers to.
- Token (words) tagging in text:
  - to connect with grammatical features
  - to (related) concepts
  - to other language(s) equivalents

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, Univesiy of Santiago de Compostela, Spain, 18-22 July 2016
Approaches

• Annotation approaches types:
  ▫ completeness of the result (i.e. how well does it capture the real-world situation) and
  ▫ commitment to the result (i.e. usability, understanding).

• For example,
  ▫ tags require little effort and result in high commitment (through the collaborative tagging), but
  ▫ they have a low completeness (one can not make complex statements about the real world, but only assign shallow tags).

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
Types of annotations

- Informal annotations,
- Formal annotations,
  - that have formally defined constituents and are thus machine-readable, and
- Ontological annotations,
  - that have formally defined constituents and use only ontological terms that are socially accepted and understood.

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSONE Training School on Keyword search in Big Linked Data, Universiy of Santiago de Compostela, Spain, 18-22 July 2016
Definition 1 (Annotation). An annotation $A$ is a tuple $(a_s, a_p, a_o, a_c)$, where $a_s$ is the subject of the annotation (the annotated data) $a_o$ is the object of the annotation (the annotating data) $a_p$ is the predicate (the annotation relation) that defines the type of relationship between $a_s$ and $a_o$, and $a_c$ is the context in which the annotation is made.

Definition 2 ( Formal annotation). A formal annotation $A_f$ is an annotation $A$, where the subject $a_s$ is a URI, the predicate $a_p$ is a URI, the object $a_o$ is a URI or a formal and the context $a_c$ is a URI.

Definition 3 (Ontological annotation). A ontological annotation $A_s$ is a formal annotation $A_f$, where the predicate $a_p$ and the context $a_c$ are an (arbitrarily complex) ontological term, and the object $a_o$ conforms\cite{10} to an ontological definition of $a_p$. 

---

Ranka Stanković. "Semantic annotation and expansion for keyword queries”. The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, Universiy of Santiago de Compostela, Spain, 18-22 July 2016
Examples

• Informal

9) above) or by movement (by adjoining an auxiliary/T-constituent above). One minor problem posed by the analysis in (69) is that it in question our earlier claim that Q is a purely verbal affix, since the doesn’t seem to be a verbal head.

• Formal

<http://papers.org/minimalism#minor>
<disagree> "that's not minor!".

• Ontological

<http://papers.org/minimalism#minor>
ibis:con
[ rdf:type ibis:Argument;
  rdf:label "that's not minor!" ].

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSSTONE Training School on Keyword search in Big Linked Data, Univesiy of Santiago de Compostela, Spain, 18-22 July 2016
How to classify annotations

• Association
  ◦ way an annotation is associated with the annotated resource - whether the annotation is embedded in the annotated resource, or references the resource externally.

• Subject granularity
  ◦ granularity of the annotation subject: e.g. is the annotation about a document, a section inside a document, a sentence, or a word?

• Representation distinction
  ◦ whether the tool distinguishes annotations about documents from annotations of the concept described in or otherwise related to the document?

• Terminology reuse (“heterogeneity”, “interoperability”)
  ◦ whether an annotation is self-confined with its own terminology, or whether an annotation uses terms from (one or more) existing ontologies, and are thus interoperable and understandable for others.

• Object type ("annotation form")
  ◦ type of annotation object: is it a literal or textual object, a structural object (including a hyperlink to another page), or an ontological object?

• Context context of the anno
  ◦ when was it made, by whom, and within what scope: the annotation could for example be temporally scoped (it is only valid in 2016) or spatially scoped (it is only valid in Spain).
  ◦ If the annotation is not about a document, then the context could also be the document the annotation is derived from.
Named entity recognition

- **Named entity recognizers** identify proper names in documents, and may also classify these proper names as to whether they designate people, places, companies, organizations, and the like.

- In the sentence:
  - *Italy’s* business world was rocked by the announcement *last Thursday* that *Mr. Verdi* would leave his job as vice-president of *Music Masters of Milan, Inc* to become operations director of *Arthur Andersen*.

- ‘Italy’ would be identified as a place, ‘last Thursday’ as a date, ‘Verdi’ as a person, ‘Music Masters of Milan, Inc’ and ‘Arthur Andersen’ as companies.

- Some would consider recognition of ‘Milan’ as a place, and identifying ‘Arthur Andersen’ as a person as an error in this context.
Named entities and coreferences

• MUC defined a coreference task as linking together multiple expressions that refer to a given entity.
• In the context of information extraction, the role of coreference annotation is to ensure that information associated with multiple mentions of an entity can be collected together.
• For instance,
  ▫ `<coref id='100'>International Business Machines </coref>`
  ▫ `<coref id='101' type='ident' ref='100'>IBM</coref>`
• The acronym **IBM** refers to the identical notion as the phrase **International Business Machines**.
Some examples of **ENAMAX** tags/3

Country name is a part of a name of an organization:

- `<ENAMEX TYPE="ORGANIZATION">Hyundai of Korea, Inc.</ENAMEX>`

Country name is not a part of a name of an organization:

- `<ENAMEX TYPE="ORGANIZATION">Hyundai, Inc.</ENAMEX>` of `<ENAMEX TYPE="LOCATION">Korea</ENAMEX>`

City name is not a part of a name of an university:

- `<ENAMEX TYPE="ORGANIZATION">University of California</ENAMEX>` in `<ENAMEX TYPE="LOCATION">Los Angeles</ENAMEX>`

Compound expressions in which place names are separated by a comma are to be tagged as separate instances of **LOCATION**

- `<ENAMEX TYPE="LOCATION">Kaohsiung</ENAMEX>`, `<ENAMEX TYPE="LOCATION">Taiwan</ENAMEX>`
Some examples of TIMEX tags

Time

• <TIMEX TYPE="TIME">twelve o'clock noon</TIMEX>
• <TIMEX TYPE="TIME">5 p.m. EST</TIMEX>

Date

• <TIMEX TYPE="DATE">January 1990</TIMEX>
• <TIMEX TYPE="DATE">fiscal 1989</TIMEX>
• the <TIMEX TYPE="DATE">autumn</TIMEX> report (?)
• <TIMEX TYPE="DATE">third quarter of 1991</TIMEX>
• <TIMEX TYPE="DATE">the fourth quarter ended Sept. 30</TIMEX>
Some examples of NUMEX tag

**monetary expressions:**

- `<NUMEX TYPE="MONEY">20 million New Pesos</NUMEX>`
- `<NUMEX TYPE="MONEY">$42.1 million</NUMEX>`
- `<NUMEX TYPE="MONEY">million-dollar</NUMEX> conferences`

**percentage**

- `<NUMEX TYPE="PERCENT">15 pct</NUMEX>`
Named Entity Categories and TEI

• One chapter of TEI (Text Encoding Initiative) guidelines is dedicated to named entities:
  ▫ **P5: Guidelines for Electronic Text Encoding and Interchange**
  ▫ Chapter 13: **Names, Dates, People, and Places**
• Elements and their attributes are described in this chapter that can be used when a special TEI module is included namesdates – without it only basic elements can be used, for instance for names those are name and rs.
Person names in TEI

- `<persName>`
  - `<surname>`
  - `<forename>`
  - `<roleName>`
  - `<addName>`
  - `<nameLink>`
  - `<genName>`

Examples:

```xml
<persName key="DUDO1">
  <roleName type="honorific" full="abb">Mme</roleName>
  <nameLink>de la</nameLink>
  <surname>Rochefoucault</surname>
</persName>

<persName>
  <forename>Charles</forename>
  <genName>II</genName>
</persName>
```
Geopolitical names in TEI

• `<placeName>`
  - `<district>`
  - `<settlement>`
  - `<region>`
  - `<country>`
  - `<bloc>`

• Examples:
  `<placeName key="LSEA1">`
  `<country type="nation">Laos</country>`,
  `<bloc type="sub-continent">Southeast Asia</bloc>`
  `<placeName>`
  `<placeName>`

  `<settlement type="city">Rochester</settlement>`,
  `<region type="state">New York</region>`
  `<placeName>`
Organization names in TEI

• <orgName> - Examples:
  ▫ About a year back, a question of considerable interest was agitated in the <orgName key="PAS1" type="voluntary"> <placeName key="PEN">Pennsylvania</placeName> Abolition Society</orgName>....
  ▫ A spokesman from <orgName type="regional"> <orgName type="acronym">IBM</orgName> <country type="acronym">UK</country> </orgName> said ...
Problems with NER

• Many referring expressions are proper names and may therefore exhibit initial capital letters in English (and many other European languages), e.g., John Smith, Thomson Corporation and Los Angeles.
• The presence of an initial capital does not guarantee that one is dealing with part of a name, since initial capitalization is also used:
  ▫ at the start of sentences,
  ▫ Variables in mathematics, chemical symbols, X-rays, ...
  ▫ Acronyms that are not named entities (FC – for football club)
  ▫ Acronyms in short messages: OMG (Oh, my God), etc.
• Also, for some named entities no initial capital letter is used, e.g. eBay.
NER system for Serbian

- Entities that are tagged belong to classes:
  - **Person names** (full names and distinguished person names) their titles, roles and functions, if present, preceding or following them;
  - **Geopolitical names** – countries and settlements – **geographic names** – water bodies and onyms.
  - **Organization names** – including names of political parties.
  - **Number expressions** – monetary, measurements, count, percentage
  - **Time expressions** – dates, times of day, periods and frequencies, absolute and relative
General resources used for the Serbian NER

• Comprehensive morphological e-dictionaries of Serbian in DELA/DELAF format:
  ▫ simple words,
  ▫ Multi-word names;

• including:
  ▫ general lexica,
  ▫ geographic names,
  ▫ personal names,
  ▫ encyclopedic knowledge (in development).

• Dictionary entries are provided with elaborate semantic markers.
Examples of Dictionary Entries

• Geographic names:
  - **Dunav,N+NProp+Top+Hyd** (Danub is a proper name, geographic notion, hydronym)
  - **Atlanski okean,N+NProp+Top+Hyd** (Atlantic Ocean)

• Geopolitical names:
  - **Madrid,N+NProp+Top+Gr** (Madrid – a proper name, city)
  - **Španija,N+NProp+Top+Dr** (Spain – a proper name, country)

• Organizations:
  - **Atinska novinska agencija,N+NProp+Org+Acr=ANA**

• Person names:
  - **Venizelos,N+NProp+Hum+Last+Cel** (a last name of a famous person)
  - **Riga od Fere,N+NProp+Hum+Last+Cel** (a full name of a famous person)
The general approach - rule-based supported by lexical resources

• Use of dictionaries

• Use of local grammars to specify the context
  ▫ For rejecting false recognitions
  ▫ For accepting false rejections

• A task: recognition and tagging of **hydronyms** (water bodies) in Serbian newspaper texts.

• Problems: hydronyms are ambiguous with:
  ▫ other geographic names: *Bosna* – a river and a region.
  ▫ personal names: *Una* – a river and a feminine name, *Sava* – a river and a masculine name
  ▫ Common nouns: *Kupa* – a river, and but als a form of a noun *kup*, a verb *kupati*...
The first solution

• We use as a text a small collection of news dealing with recent floods in Serbia in 2014 named *Poplave* (~10,000 simple words)
• For retrieving names of water bodies we use a pattern:
  ▫ `<N+NProp+Top+Hyd>`
• All names of water bodies (recorded in e-dictionaries) but also a number of false recognitions:
  ▫ *Oko* – a preposition ’around’ and a form of a name *Oka*
• 89 matches / 7 false matches
The improvement

• Use local grammar that take into consideration only simple word hydronym names that are not ambiguous with other proper or common names.

• Recognize MWU hydronym names in dictionaries (usually they are not ambiguous)

• Use context that have (river, lake, on the bank of..., hydropower on...)

• Use context that have („has flooded“...)

• Check context –if it is in list with already recognized water body names.
The last improvement

- We try to retrieve some more entries – even those uppercase unknown words but with an obligatory key word following or preceding it
  - 76 matches in a collection Poplave / no false recognitions / 82 correct hydronym names
  - differences from a previous recognition
  - Example: Tulovska reka, (reaka) Lugomir

Прелиминарна техноекономска студија услова и могућности експлоатације лежишта угља Черевић Нови Сад, Фрушка Гора, Рударски институт Београд. Ову студију је требало урадити на основу података из Елабората о прорачуну резерви угља Ц2 категорије између Беочина и Баноштора. Институт за грађевинарство Суботица, Миливој Макар, дипл. инж. руд. Беочин. Истраживано подручје налази се на северним падинама Фрушке Горе, између Беочина на истоку и Баноштора на западу. Угљени слојеви се даље настављају према западу до Корушке. Студија о хидрогеолошким истраживањима у зони између Параћина и Главице у циљу отварања новог изворишта (I фаза). Београд Др Војислав Томић, доцент, Невен Крешић, дипл.инж.

Прелиминарна техноекономска студија услова и могућности експлоатације лежишта угља Черевић Нови Сад, Фрушка Гора, Рударски институт Београд. Ову студију је требало урадити на основу података из Елабората о прорачуну резерви угља Ц2 категорије између Беочина и Баноштора. Институт за грађевинарство Суботица, Миливој Макар, дипл. инж. руд. Беочин. Истраживано подручје налази се на северним падинама Фрушке Горе, између Беочина на истоку и Баноштора на западу. Угљени слојеви се даље настављају према западу до Корушке. Студија о хидрогеолошким истраживањима у зони између Параћина и Главице у циљу отварања новог изворишта (I фаза). Београд Др Војислав Томић, доцент, Невен Крешић, дипл.инж.
Tools

• Manual annotation tools: Annotea, OntoMat, COHSE, WebAno,…
• (semi-)automatic annotation GATE, Unitex, NooJ,…
• Statistical tagging: treetagger, Stanford POS tagger,…
• Statistical natural language processing and corpus-based computational linguistics:
  ▫ An annotated list of resources
    • Tools: Machine Translation, POS Taggers, NP chunking, Sequence models, Parsers, Semantic Parsers/SRL, NER, Coreference, Language models, Concordances, Summarization,…
    • Corpora: Large collections, Particular languages, Treebanks, Discourse, WSD, Literature, Acquisition
    • SGML/XML
    • Dictionaries, Lexical/morphological resources
    • Courses, Syllabi, and other Educational Resources
    • Mailing lists
    • Other stuff on the Web: General, IR, IE/Wrappers, People, Societies
The semantic annotation of texts consists in extracting semantic relations between domain relevant terms in texts. Several studies address the problem of capturing complex relations from texts. They combine statistical and linguistic analyses. In the semantic annotation generation, the aim is to identify existing relations, belonging to the domain ontology, within instances in texts and to complete them with the description of the domain concepts related by these identified relations.

Semantic Annotation of Texts with RDF Graph Contexts

- Mapping between grammatical elements of each sentence in the analysed text and the corresponding entities in the dedicated-domain ontology.
  - the detection of relations described in a domain ontology,
  - the detection of terms linked by the identified relations based on term linguistic roles (subject, object, etc.) in the sentence, and
  - the generation of a corresponding annotation of the analysed text.
- Distinguishing between the ontological level and the instance level when linking a term in the text to the ontology: a term is identified to an instance of a concept rather than to the concept itself;
- Enriching the extracted instances of conceptual relations with contextual knowledge.
- Corese semantic search engine: RDF graph-based knowledge representation, SPARQL and RDF contextual metadata (contexts).

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, Univesiy of Santiago de Compostela, Spain, 18-22 July 2016
NLP-Driven Semantic Annotation of Texts

Extraction of relations from texts (RASP for English)

Mapping of grammatical constituents to RDF triples (RASP output » RDF triples)

---

The L1 luggage compartment contains 100cc.

<table>
<thead>
<tr>
<th>RASP syntactic tree analysis</th>
<th>RDF annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(&quot;S&quot; (&quot;NP&quot; (&quot;NP&quot; &quot;The&quot; &quot;L1&quot;) &quot;luggage&quot; &quot;compartment&quot;) (&quot;VP&quot; &quot;contain::s&quot; (&quot;NP&quot; &quot;100cc&quot;)))</td>
<td>&lt;spro:Luggage_compartment rdf:about=&quot;#L1&quot;&gt; <a href="">spro:contain</a> &lt;spro:Capacity rdf:about=&quot;#100cc&quot; /&gt; &lt;/spro:contain&gt; &lt;/spro:Luggage_compartment&gt;</td>
</tr>
</tbody>
</table>
NLP-Driven Semantic Annotation of Texts

- S – V – O (sentence in active form)
  - O – V – S (sentence in passive form)
  - subordinate phrases “independent” from the main sentence and rhetorical relations
  - ambiguous subject/object constituents

- constitutes are major problem and may lead to a deadlocks when querying with SPARQL
  - “RDF graph context” with recursive capability for more expressive structures

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEystONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
<table>
<thead>
<tr>
<th>CategoryID</th>
<th>SubCategID</th>
<th>Caption in English</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>person</td>
<td>pers</td>
<td>Who is?</td>
<td>20</td>
</tr>
<tr>
<td>person</td>
<td>function</td>
<td>What does he/she do?</td>
<td>5</td>
</tr>
<tr>
<td>organization</td>
<td>organization</td>
<td>What is it?</td>
<td>25</td>
</tr>
<tr>
<td>location</td>
<td>location</td>
<td>At which place/where?</td>
<td>42</td>
</tr>
<tr>
<td>count</td>
<td>measure</td>
<td>How much?</td>
<td>7</td>
</tr>
<tr>
<td>count</td>
<td>money</td>
<td>How much money?</td>
<td>0</td>
</tr>
<tr>
<td>count</td>
<td>amount</td>
<td>How much of what?</td>
<td>4</td>
</tr>
<tr>
<td>count</td>
<td>percent</td>
<td>What percentage?</td>
<td>0</td>
</tr>
<tr>
<td>time</td>
<td>date</td>
<td>Which day?</td>
<td>10</td>
</tr>
<tr>
<td>time</td>
<td>hour</td>
<td>What time?</td>
<td>5</td>
</tr>
<tr>
<td>time</td>
<td>duration</td>
<td>How long does it take?</td>
<td>0</td>
</tr>
<tr>
<td>time</td>
<td>frequency</td>
<td>How many times repeated</td>
<td>0</td>
</tr>
</tbody>
</table>
Extending Full Text Search Engine for Mathematical Content

- Index and search for mathematical content on the WWW using full text search engine
- Linearization, transformation rules, generalisation rules and ordering algorithm simplify the complex and highly symbolic mathematical structures into linear structures with well-defined symbols

1. Partial evaluation: $7 + a + 5 \xrightarrow{\text{converted to}} 12 + a$
2. Approximate numerical constants: $5.82 \div 6$
3. Remove brackets using distributivity: $a \ast (b + c) \rightarrow a \ast b + a \ast c$
4. Multiply tokens: $\frac{a+b}{2} \ast \Pi \rightarrow \frac{\Pi a + \Pi b}{2}$
5. Assign each numerator its own denominator: $\frac{\Pi a + \Pi b}{2} \rightarrow \frac{\Pi a}{2} + \frac{\Pi b}{2}$
6. Replace constants with const symbol: $74 + a^2 + b^2 \rightarrow \text{const} + a^{\text{const}} + b^{\text{const}}$
7. Replace unknown constants, variables with id symbol: $a^2 - b^2 + 2bc \rightarrow \text{id}_1^2 - \text{id}_2^2 + 2\text{id}_1\text{id}_2$
   or $\rightarrow \text{id}_1^2 - \text{id}_2^2 + 2\text{id}_1\text{id}_2 \ldots$

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
Extending Full Text Search Engine for Mathematical Content

• Searching phase is the only user interactive phase of a search engine
• User enters a query (in LaTex) which is executed and the results are displayed.
• This includes several steps:
  ▫ 1) query parsing,
  ▫ 2) mapping query operators to supported internal constructs,
  ▫ 3) finding all words/phrases from the query,
  ▫ 4) evaluating the logic of the query and collecting suitable documents,
  ▫ 5) sorting them according to their rank,
  ▫ 6) displaying the result list.
• The mathematical extension is part of 1), 2) and 6).

Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEYSTOLE Training School on Keyword search in Big Linked Data, Univesiy of Santiago de Compostela, Spain, 18-22 July 2016
Ranka Stanković. "Semantic annotation and expansion for keyword queries". The 2nd KEystONE Training School on Keyword search in Big Linked Data, University of Santiago de Compostela, Spain, 18-22 July 2016
Figure 3: Math-aware search in MlaS
<table>
<thead>
<tr>
<th>System</th>
<th>Input documents</th>
<th>Internal representation</th>
<th>Approach</th>
<th>$\alpha$-eq.</th>
<th>Query language</th>
<th>Queries</th>
<th>Indexing core</th>
</tr>
</thead>
<tbody>
<tr>
<td>MathDex</td>
<td>HTML, TeX/LaTeX, Word, PDF</td>
<td>Presentation MathML (as strings)</td>
<td>syntactic</td>
<td>$\times$</td>
<td>?</td>
<td>text, math</td>
<td>Apache Lucene</td>
</tr>
<tr>
<td>LeActiveMath</td>
<td>OMDoc, OpenMath</td>
<td>OpenMath (as string)</td>
<td>syntactic</td>
<td>$\times$</td>
<td>OpenMath (palette editor)</td>
<td>text, math, mixed</td>
<td>Apache Lucene</td>
</tr>
<tr>
<td>\LaTeXSearch</td>
<td>\LaTeX</td>
<td>\LaTeX (as string)</td>
<td>syntactic</td>
<td>$\times$</td>
<td>\LaTeX</td>
<td>titles, math, DOI</td>
<td>?</td>
</tr>
<tr>
<td>MathWeb Search</td>
<td>Presentation MathML, Content MathML, OpenMath</td>
<td>Content MathML, OpenMath (substitution trees)</td>
<td>semantic</td>
<td>$\checkmark$</td>
<td>QMath, \LaTeX, Mathematica, Maxima, Maple, Yacas styles (palette editor)</td>
<td>text, math, mixed</td>
<td>Apache Lucene (for text only)</td>
</tr>
<tr>
<td>EgoMath</td>
<td>Presentation MathML, Content MathML, PDF</td>
<td>Presentation MathML trees (as strings)</td>
<td>mixed</td>
<td>$\checkmark$</td>
<td>\LaTeX</td>
<td>text, math</td>
<td>EgoThor</td>
</tr>
<tr>
<td>MIAaS</td>
<td>any (well-formed) MathML</td>
<td>Canonical Presentation MathML trees (as compacted strings)</td>
<td>math tree similarity/normalization</td>
<td>$\checkmark$</td>
<td>AMS-\LaTeX or MathML</td>
<td>text, math, mixed</td>
<td>Apache Lucene/Solr</td>
</tr>
</tbody>
</table>

- EuDML – European digital mathematics library https://eudml.org/
- WebMIaS is a web interface for Math Indexer and Searcher (MIaS) math aware searching engine, https://mir.fi.muni.cz/webmias-demo/
- WolframAlfa
"Far too many algorithms in use today are being used as weapons against populations, whether they are consumers, workers, prisoners, or teachers. I'll talk about a few which I consider the worst kind - and which I call Weapons of Math Destruction - namely, those that are opaque, widespread, and powerful enough to cause tremendous destruction through feedback loops. I will also discuss suggestions for data scientists, policy makers, and the public for how to combat them."
For more details see

Thank you for your attention

Hvala na pažnji
Хвала на пажњи