# Semantic search

## Creating and querying semantic index

Jan Kouřil

Natural Language Processing Group
Faculty of Information Technology
Brno University of Technology

May 16, 2013

Supervisor: Doc. RNDr. Pavel Smrž, Ph.D.

# Outline

Motivation and Objectives
Semantic index
Features
Future Work

Introduction
The Decipher project
Goals of work

# Outline

Motivation and Objectives
Semantic index
Features
Future Work

Introduction
The Decipher project
Goals of work

# Motivation

- Search accuracy can be improved by understanding searcher's intent
- Contextual meaning of lexical units can be resolved easier for specific domain
- Knowledge base for any domain may be used by many different users
- Integrating user feedback is a way to improve search
- Rich query language may help a lot

Motivation and Objectives
Semantic index
Features
Future Work

Introduction
The Decipher project
Goals of work

## Techniques used

- Named entity recognition
- Co-reference resolution
- Finite automata for lexical purposes
- Classification algorithms for document analysis
- Document indexing

Motivation and Objectives
Semantic index
Features
Future Work

Introduction
The Decipher project
Goals of work

## Decipher overview

- Digital Environment for Cultural Interfaces
- Promoting Heritage
- Education and Research

- Storyscope front-end
- SEC back-end

Motivation and Objectives
Semantic index
Features
Future Work

Introduction
The Decipher project
Goals of work

# Storyscope

Dashboard  Content  Structure  Appearance  People  Modules  Configuration  Reports  Help

Hello admin   Log out

Add content  Find content

Edit shortcu

## Search

"world war*"                                                                    Search

2466 results                          Context:  None                      ▼

▼ Facets

☐ The Nation's War of the World                          Do not show

http://www.vads.ac.uk/farge.php?sos=177&uid=29595
Edith Cavall (1865-1915) was a Red Cross nurse in Belgium, executed by the Germans during
the First **World War**.

☐ First World War scene                                  Do not show

http://artsalesindex.artinfo.com/asi/lots/2848452
First **World War** scene by Francois-Charles Baude in Works of Art, Antique &
Decorative Furniture on June 06, 2000 at the null null sale B00230, lot 657

☐ First World War battle scene                           Do not show

http://artsalesindex.artinfo.com/asi/lots/377510
First **World War** battle scene by Douglas S Andrews in Modern
Paintings & British & Continental Watercolours on May 22, 2001
at the null null sale 28558, lot 529

☐ After World War I and II                               Do not show

http://artsalesindex.artinfo.com/asi/lots/560051
After **World War** I and II by Marcello Andreozzi in Tuscan Paintings & 19th Century
Graphic Arts on March 30, 2001 at the null null sale No. 100, lot 423

**Where found:**
☐ External (2455)
☐ Internal (11)

**Has image:**
☐ True (2203)
☐ False (263)

**Source dataset:**
☐ Vads (850)
☐ Artsalesindex (752)
☐ Freebase (402)
☐ Artnet (310)
☐ Europeana (95)
☐ NGI (39)
☐ IMMA (5)
☐ Wikipedia (2)
☐ Guardian (1)

**Event:**
☐ World War II (30)
☐ World War I (25)
☐ Franco Prussian War (11)
☐ American Civil War (11)
☐ Spanish Civil War (9)
☐ Battle of Jutland (9)
☐ Cold War (8)
☐ Western Front (7)
☐ Vietnam War (7)
☐ Eastern Front (5)

**Artist:**
☐ Johannes Wohnseifer (8)
☐ Salvador Dalí (7)
☐ M C Escher (6)
☐ Damien Hirst (6)
☐ Pablo Picasso (5)
☐ Cai Yuan (4)
☐ Andre Carl (4)
☐ Bhatt Jyoti (3)
☐ Bechtle Robert (3)

Motivation and Objectives
Semantic index
Features
Future Work

Introduction
The Decipher project
Goals of work

# Semantic Enrichment Component

- Query language
- Faceted search
- Event detection
- Handling index
- Content classification

Motivation and Objectives
Semantic index
Features
Future Work

Introduction
The Decipher project
Goals of work

## Dossier creation

1. Search artworks with query language
2. Add artworks to dossier
3. Extract events and form plots
4. Create narratives
5. Publish dossier

Motivation and Objectives
**Semantic index**
Features
Future Work

Index structure
Query language

# Outline

Motivation and Objectives
**Semantic index**
Features
Future Work

Index structure
Query language

## Elasticsearch

- Based on Apache Lucene
- RestFul server
- Tools for analysis (tokenizers, stemmers)
- Key-value store
- Multicast cluster

Motivation and Objectives
**Semantic index**
Features
Future Work

Index structure
Query language

# Indexed types

- Dossiers
- Object stories
- Artworks
- References
- Events
- Plot elements
- Plots

Motivation and Objectives
**Semantic index**
Features
Future Work

Index structure
Query language

# Example 1

Mona Lisa; dataset: wikipedia; intended_audience: youths

Context: None

Do not show

☐ Mona Lisa **Mona Lisa**



http://simple.wikipedia.org/wiki/Mona_Lisa

**Mona Lisa** (also known as La Gioconda or La Joconde) is a 16th-century portrait painted in oil by Leonardo da Vinci during the Renaissance in Florence, Italy. As of 2011, the work is owned by the Government of France and is shown at the Musée du Louvre in Paris under the title Portrait of **Lisa** Gherardini, wife of Francesco del Giocondo. Many people think **Mona Lisa**'s smile is mysterious. It is so often studied, recognized, and copied that many people say it is the most famous painting in the world.The Louvre says that about 80 percent of its visitors come to see the painting of **Mona Lisa**.

Vasari, who was Leonardo's first biographer (a person who writes about the life of another person), thought the painting was of a person named **Lisa** Gherardini. Speculation over the painting's model was solved in 2008 by Dr Armin Schlechter, a manuscript expert. Notes discovered in Heidelberg University Library which were written by Agostino Vespucci, a Florentine city official, reinforced Vasari's earlier identification of the model. **Lisa** was the wife of Francesco del Giocondo a wealthy silk merchant, who lived in Florence. Although the sitter has traditionally been identified as **Lisa** del Giocondo, a lack of definitive evidence has long fueled alternative theories, including Leonardo's mother Caterina in a distant memory as the painting's

Motivation and Objectives
**Semantic index**
Features
Future Work

Index structure
**Query language**

# Example 2

Motivation and Objectives
**Semantic index**
Features
Future Work

Index structure
**Query language**

# Example 3

person.style: impressionism OR postimpressionism; mentions.place: not in Europe

Context: None

Do not show

**Paul Gauguin**

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

Gauguin was born in Paris, France, to journalist Clovis Gauguin and Alina Maria Chazal, daughter of the proto-socialist leader Flora Tristan, a feminist precursor whose father was part of an influential Peruvian family. In 1850 [4] the family left Paris for that country, motivated by the political climate of the period.[citation needed] Clovis died on the voyage, leaving eighteen-month-old Paul, his mother, and sister, to fend for themselves. They lived for four years in Lima with Paul's uncle and his family. The imagery of Peru would later influence Gauguin in his art. It was in Lima that Gauguin encountered his first art. His mother admired Pre-Columbian pottery, collecting Inca pots that some colonists dismissed as barbaric. "One of Gauguin's few early memories of his mother was of her wearing the traditional costume of Lima, one eye peeping from behind her manteau, the mysterious one-eye veil that all women in Lima went out in. [...] He was always drawn to women with a 'traditional' look. This must have been the first of the colourful female costumes that were to haunt his imagination."[5] At the age of seven, Gauguin and his family returned to France, moving to Orléans to live with his grandfather. The Gauguins came originally from the area and were market gardeners and greengrocers: gauguin means 'walnut-grower'. His father had broken with family tradition to become a journalist in Paris. Gauguin soon learned French, though his first and preferred language remained Peruvian Spanish, and he excelled in his studies. After attending a

Motivation and Objectives
**Semantic index**
Features
Future Work

Index structure
Query language

## Implementation

- Implemented in Python
- BoolQuery with list of constraints
- Named Entity Extraction
- Information hierarchy expansion
- Query expansion

Motivation and Objectives
Semantic index
**Features**
Future Work

Autocomplete
Classification
User feedback
Facets
Similarity search

# Outline

Motivation and Objectives
Semantic index
**Features**
Future Work

Autocomplete
Classification
User feedback
Facets
Similarity search

## Features

- Query autocomplete
- Facets
- Similarity search
- User feedback
- Classification

Motivation and Objectives
Semantic index
Features
Future Work

Autocomplete
Classification
User feedback
Facets
Similarity search

## Autocomplete

- Artist names
- Artworks
- Event activities
- Genres
- Artistic styles
- Locations
- Event activities
- Others...

Motivation and Objectives
Semantic index
Features
Future Work

Autocomplete
Classification
User feedback
Facets
Similarity search

## Autocomplete implementation

- Wildcard queries
- Index lookup
- Geonames database
- Autocomplete with constraints

Motivation and Objectives
Semantic index
Features
Future Work

Autocomplete
Classification
User feedback
Facets
Similarity search

## Classification

- Choose categories to classify
- Select features
- Choose algorithm
- Take facet

# User feedback

- Positive and negative feedback
- User-specific
- Improve average ranking

Motivation and Objectives
Semantic index
**Features**
Future Work

Autocomplete
Classification
User feedback
**Facets**
Similarity search

# Facets

- From datasets
- Classified facets
- Events
- Dynamic facets

Motivation and Objectives
Semantic index
**Features**
Future Work

Autocomplete
Classification
User feedback
Facets
**Similarity search**

## Techniques

- MoreLikeThis
- MoreLikeThese
- TF-IDF query terms
- Query constraints
- LDA

Motivation and Objectives
Semantic index
Features
Future Work

Autocomplete
Classification
User feedback
Facets
Similarity search

## LDA experiments

- The most similar topic
- Overall topic similarity
- The most overlapping topics
- Combined measures
- Synonymy replacement

Motivation and Objectives
Semantic index
Features
Future Work

Payloads
Knowledge-based co-reference

# Outline

Motivation and Objectives
Semantic index
Features
Future Work

Payloads
Knowledge-based co-reference

## Future development

- Improve metadata querying
- Implement payload queries
- Improve query language
- Knowledge-based co-reference

Motivation and Objectives
Semantic index
Features
Future Work

Payloads
Knowledge-based co-reference

## Payloads

| Position | Word | NER |
|:---:|:---:|:---:|
| 1 | The | Person |
| 2 | man | Person |
| 3 | went | Activity |
| 4 | to | Activity |
| 5 | the | Location |
| 6 | store | Location |

Motivation and Objectives
Semantic index
Features
Future Work

Payloads
Knowledge-based co-reference

# Knowledge-based co-reference

Examples

- Later, he moved to the south of France and was influenced by the strong sunlight he found there. His work grew brighter in color...

- David and The Last Judgement are perhaps the best known artworks from Italian renaissance artist Michelangelo. The sculpture is located in Galleria dell'Accademia, Florence.

Extractions

- An artist (Vincent van Gogh) was influenced by an event (moving to south France)

- Sculpture of David is located in Florence.