Improving web search with FCA

Radim BELOHLAVEK
Jan OUTRATA

Dept. Systems Science and Industrial Engineering
Watson School of Engineering and Applied Science
Binghamton University – SUNY, NY, USA

Dept. Computer Science
Faculty of Science
Palacký University, Olomouc, Czech Republic
Information Retrieval × Formal Concept Analysis

web search = mining web retrieval results, part of web mining

**Information Retrieval (IR)** = retrieval of required information from textual unstructured or semistructured data (example: search by keywords, retrieval of documents), iterative and interactive process (mining):
- submitting query,
- looking at the data returned,
- submitting a refined query until appropriate data are found.

**Formal Concept Analysis (FCA)** = method of analysis of tabular data, extracting a hierarchically ordered collection of clusters:
- (input) tabular data = objects described by attributes,
- (output) clusters = objects having common attributes (and vice versa),
- used for data mining, knowledge discovery, preprocessing data, clustering and classification (conceptual clustering) etc.
rationale behind using FCA in IR and document mining:

- current search engines (e.g. Google, Yahoo, etc.) provide a ranked list of retrieved documents, i.e. a “simplistic” linear view on retrieved information, without the possibility to inspect related documents at the same time,

- FCA enables structured (or categorized) view of retrieved information with contextual information,

- user is supplied with a (part of a) conceptual hierarchy of retrieved documents and he or she can browse the hierarchy to find required information more quickly,

- new type of information can be mined: most common/uncommon subjects, which subjects imply or are implied by other subjects, novel subject associations etc. → Conceptual Knowledge Processing
Formal Concept Analysis (FCA)

FCA = method of analysis of **tabular data** (Wille, TU Darmstadt, 1982)
- alternatively called: concept data analysis, concept lattices, ...
- used for **data mining** and **knowledge discovery**

**input:**

<table>
<thead>
<tr>
<th>( I )</th>
<th>( y_1 )</th>
<th>( y_2 )</th>
<th>( y_3 )</th>
<th>( X = { x_1, x_2, \ldots } )</th>
<th>set of <strong>objects</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>( x_1 )</td>
<td>( X )</td>
<td>( X )</td>
<td>( X )</td>
<td>( Y = { y_1, y_2, \ldots } )</td>
<td>set of <strong>attributes</strong></td>
</tr>
<tr>
<td>( x_2 )</td>
<td>( X )</td>
<td>( X )</td>
<td>( )</td>
<td>( I \subseteq X \times Y )</td>
<td>relation <strong>to have</strong></td>
</tr>
<tr>
<td>( x_3 )</td>
<td>( X )</td>
<td>( X )</td>
<td>( )</td>
<td>( \langle x, y \rangle \in I )</td>
<td><strong>object</strong> ( x ) <strong>has attribute</strong> ( y )</td>
</tr>
</tbody>
</table>

**output**
- **concept lattice** (hierarchically ordered set of clusters – **formal concepts**)
- **attribute implications** (particular attribute dependencies)
FCA basics

\[ l \\
\begin{array}{ccc}
 y_1 & y_2 & y_3 \\
x_1 & X & X & X \\
x_2 & X & X \\
x_3 & X & X \\
\end{array} \Rightarrow \text{induced operators}
\]

\[ \ldots \text{mappings } \uparrow : 2^X \rightarrow 2^Y, \downarrow : 2^Y \rightarrow 2^X: \]

\[ A^{\uparrow} = \{ y \in Y \mid \forall x \in A : (x, y) \in l \} \]

\[ B^{\downarrow} = \{ x \in X \mid \forall y \in B : (x, y) \in l \} \]

\[ A \subseteq X \mapsto A^{\uparrow} \ldots \text{attributes common to all objects from } A \]

\[ \{x_1, x_2\}^{\uparrow} = \{y_1, y_3\} \]

\[ B \subseteq Y \mapsto B^{\downarrow} \ldots \text{objects sharing all attributes from } B \]

\[ \{y_1, y_2\}^{\downarrow} = \{x_1\} \]

(Birkhoff 1940s, Ore, Barbut & Monjardet, Wille 1982)

Definition (formal concept = fixed point of \( \uparrow, \downarrow \))

Formal concept in data is a pair \( \langle A, B \rangle \) s.t.

\[ A^{\uparrow} = B \text{ and } B^{\downarrow} = A. \]

formal concepts \( \approx \) all potentially interesting clusters in data
FCA basics

Definition (concept lattice = formal concepts + concept hierarchy)

Concept lattice (Galois lattice) of $\langle X, Y, I \rangle$ is the set

$$B(X, Y, I) = \{(A, B) \mid A^{\uparrow} = B, B^{\downarrow} = A\}$$

of all formal concepts PLUS concept hierarchy $\leq$ defined by

$$(A_1, B_1) \leq (A_2, B_2) \iff A_1 \subseteq A_2 \text{ (iff } B_2 \subseteq B_1).$$

FCA ... inspired by Port-Royal (traditional) approach to concepts:

- concept (according to Port-Royal) := extent $A +$ intent $B$
  - extent = objects covered by concept
  - intent = attributes covered by concept
- example: DOG (data = animals $\times$ animals' attributes)
  - extent = collection of all dogs (beagle, collie, poodle, ...)
  - intent = all dogs' attributes (barks, has four limbs, has tail, ...)
- conceptual hierarchy $\leq$ ... subconcept/superconcept relation
  - concept1=(extent1,intent1) $\leq$ concept2=(extent2,intent2)
    $\iff$ extent1 $\subseteq$ extent2 (iff intent1 $\supseteq$ intent2)
  - example: BEAGLE $\leq$ DOG $\leq$ MAMMAL $\leq$ ANIMAL
Formal concepts = maximal rectangles in data

Theorem (formal concepts = maximal rectangles)

\[ \langle A, B \rangle \text{ is a formal concept IFF } \langle A, B \rangle \text{ is a maximal rectangle.} \]

formal concepts (= maximal rectangles)

\[ (A_1, B_1) = (\{x_1, x_2, x_3, x_4\}, \{y_3, y_4\}) \]

\[ (A_2, B_2) = (\{x_1, x_3, x_4\}, \{y_2, y_3, y_4\}) \]

\[ (A_3, B_3) = (\{x_1, x_2\}, \{y_1, y_3, y_4\}) \]
Literature on FCA

- books:

- conferences: ICFCA (Int. Conf. on Formal Concept Analysis), CLA (Concept Lattices and Their Applications), ICCS (Int. Conf. on Conceptual Structures)

- web: useful resources and links at http://www.upriss.org.uk/fca/fca.html (“FCA Homepage”)

- state of the art:
  - theoretical foundations,
  - algorithms,
  - increasingly popular applications (information retrieval, software engineering, . . . ),
  - interaction with other methods of data analysis (preprocessing),
  - software available.
Selected applications of FCA

- software engineering
- association rule mining – closed frequent itemsets instead of frequent itemsets $\Rightarrow$ non-redundant association rules (much less than by usual approach)
- (Boolean) factor analysis – factors = selected formal concepts
  ... “new attributes”
- **information retrieval**, knowledge extraction – structured view on data
- machine learning (decision making), clustering and classification – preprocessing input data
- ...

see the slides “Relational Data Analysis: Applications of Formal Concept Analysis (FCA)”
FCA in Information Retrieval

pioneering work of R. Godin; C. Carpineto, G. Romano; elaborated by P. Eklund, J. Ducrou

main ideas:

- formal context = documents (objects) + index terms (attributes)
- (query/search) formal concept = (query) terms (intent) + retrieved documents (extent)
- query concept neighbors = minimal conjunctive refinements (specialization), enlargements (generalization) and alterations (categorization) of the query
Improving search engines with FCA

basic ideas:

- forwarding user query to a (web) search engine (Google, Yahoo etc., in a format such as SOAP), receiving ranked results (typically in XML format),
- parsing (first) results, indexing the document/snippet/title terms, optionally ranking the results,
- establishing formal context (possibly with attribute ordering = thesaurus),
- computing (part of the) concept lattice of the results, optionally ranking the results, displaying it to the user and
- enabling the user to appropriately modify the query by navigating through the lattice of the results (around the query concept)

more detailed treatment in Carpineto C., Romano G.: Concept Data Analysis. Wiley, 2004 (Chap. 3, 4).
Improving search engines with FCA

indexing the document terms (studied in Information Retrieval):

- text segmentation
- word stemming – using a rule-based stemmer (e.g. Porter’s) or a lexical knowledge base
- stop wording
- word weighting – crucial, “term frequency-inverse document frequency” (tf-idf) scheme implemented (most often) by a vector space model with a suitable weighting function, for web documents also URL, title, links etc.
- word selection – removing terms with low weight
- document ranking

can be seen as a feature/attribute selection problem from data mining
Improving search engines with FCA

document ranking (concept-lattice based ranking):
  - similar to hierarchical clustering-based ranking
  - **conceptual distance** between query/search concept and other document concepts in concept lattice instead of heuristic metric
  - overcomes the vocabulary problem (word mismatch) seen in best-match ranking (used by current search engines)

possible difficulties:
  - computational constraints → **computing part of the concept lattice around the query concept** = neighbor-like algorithms
  - effective concept lattice visualization → **show query concept neighborhood only** (focus+context techniques, tree below query concept)

existing (prototype) systems: CREDO, FooCA, SearchSleuth
CREDO

- system for Conceptual REorganization of DOcuments, developed by Carpineto and Romano at Fondazione Ugo Bordoni, Italy
- displays the **upper part** (two levels from the top element) of the iceberg **concept lattice** (adding terms down the lattice), in the form of a tree
- enables “offline” navigation in concepts, narrowing the scope of the search
- Carpineto C., Romano G.: Exploiting the Potential of Concept Lattices for Information Retrieval with CREDO. J. Universal Computer Science 10(8)(2004), 985–1013
- search tool available at http://credo.fub.it
- mobile version CREDINO, http://credino.dimi.uniud.it

**illustration:**
- search for “dwarf” (ambiguous term), “phoenix”,
- compare the results obtained by Credo vs. Google or Yahoo
A Topological Framework for Formal Concept Analysis

on formal concept analysis in view of the rich content of algebraic topology. 1. Introduction. The idea of formal. 1 ... Formal Concepts and Concept Lattices ...
www.l.aae.cuhk.edu.hk/~cpkwong/fca01.pdf

Formal Concept Analysis to Learn from the Sisyphus III Material
... illustrate the ideas behind Formal Concept Analysis a brief introduction of its ... For formal concepts a natural subconcept/superconcept relationship can then be ...
hki.cpsc.ucalgary.ca/KAW/KAW98/erdmann

Formal Concept Analysis And Delayed Greedy algorithm for Min-Test-Suite
Introduction of Concept Analysis. • Formal context • Common ... Define the strongest concepts as the elements in the lattice which is next to bottom. ...
www.cs.arizona.edu/classes/cs610/fa006/concept1.pdf

Introduction to FCA
Introduction to FCA. Formal Concept Analysis (FCA) is based on mathematical order theory and is a ... groups are called concepts which can be represented ...
segwald.im.unibe.ch/8080/SCG/609

Formal Concept Analysis with ConImp: Introduction to the Basic Features
In the following we try to explain the basic concepts of formal concept analysis, ... belongs to the essential basic concepts of formal concept analysis: ...
www.mathematik.tu-darmstadt.de/~burmeister/ConImpIntro.pdf
FooCA

- FCA + Google, developed by Bjoern Koester at Webstrategy GmbH, Darmstadt and TU Dresden, Germany
- presents search results directly in a form of **formal context** (documents × terms), additionally represented by labelled **Hasse diagram of the concept lattice** (clicking in the table or on the diagram nodes opens a browser window with URLs)
- “online” navigation in concepts – adding or removing attributes triggers new search and concept hierarchy formation
- search tool at [http://fooca.webstrategy.de](http://fooca.webstrategy.de) – requires registration
Improving web search with FCA

Your FooCA search for **Formal Concept Analysis** brought these results:

<table>
<thead>
<tr>
<th>G/M</th>
<th>X (10) analysis</th>
<th>X (6) concept</th>
<th>X (3) method</th>
<th>X (3) data</th>
<th>X (1) conference held</th>
<th>X (2) lattices</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

6 out of 88 attributes selected.  

- [Export the Formal Context](http://www.kuocentral.org/resources/fca.html)  
- [FlashLattice](http://www.kuocentral.org/resources/fca.html)  

About FooCA and Terms of Use  

FooCA is powered by Yahoo! Search
**SearchSleuth**

- developed by Peter Eklund and Jon Ducrou within KVO (Knowledge, Visualization and Ordering), University of Wollongong, Australia, following ImageSleuth in the conceptual neighborhood paradigm
- displays the **neighbors and siblings of the query/search concept** (direct query generalization, specialization and categorization), in the form of text labels (links) of terms/attributes determining the concepts
- “online” navigation, multiple searches per query – for neighbors of query concept, to expand the formal context
- search tool available at http://www.kvocentral.org/software/searchsleuth.html

**illustration:**
- search for “dwarf” (ambiguous term), “phoenix”,
- compare the results obtained by SearchSleuth vs. Google or Yahoo
SearchSleuth

1. Formal Concept Analysis Homepage
Formal Concept Analysis is a method of conceptual knowledge representation and data analysis. ... Christian Lindig's Concepts, (in C, older version: TkConcept? ...
www.upriss.org.uk/fca/fca.html

2. Formal concept analysis - Wikipedia, the free encyclopedia
... example concepts satisfy the formal definitions; the ... describing formal concept analysis for computer scientists. A Formal Concept Analysis Homepage ...
en.wikipedia.org/wiki/Formal_concept_analysis

3. Formal Concept Analysis
Formal Concept Analysis is a branch of applied mathematics. ... Several books on Formal Concept Analysis have appeared, among them the first ...
www.math.tu-dresden.de/~ganter/fba.html

4. Formal Concept Analysis
Formal Concept Analysis (FCA) is a method mainly used for the analysis ... into units which are formal abstractions of concepts of human thought, allowing ...
www.cs.cmu.edu/afs/cs.cmu.edu/project/ijar/pub/volume24/cimiano05a-html/node3...

5. Linguistic Applications of Formal Concept Analysis
scribes the role that formal concept analysis can play in the automated or ... Associative and Formal Concepts. In: Priss; Corbett; Angelova (eds.), Con ...
www.upriss.org.uk/papers/fcaic03.pdf
Further usage of the approach

(existing) usage besides web search:

- digital library search (Virtual Museum of the Pacific, requires registration),
- scientific (biology, medicine, ...) or social records mining,
- annotated multimedia archive search (ImageSleuth, DVDSleuth),
- email message search (MailSleuth),
- software documentation search,
- ... searching any other database of interest.

Possible usage/improvements:

- other difficult IR tasks, e.g. natural language processing
- integration with IR techniques