# Improving Conceptual Search Results Reorganization Using Term-Concept Mappings Retrieved from Wikipedia

C. Săcărea[1], R. Meza[1], M. Cimpoi[1]
[1]Babeş-Bolyai University, Cluj-Napoca, csacarea@math.ubbcluj.ro, mezaradu@yahoo.com, cm20294@scs.ubbcluj.ro

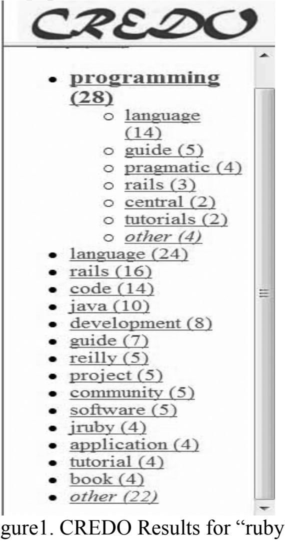
***Abstract*-This paper describes a way of improving search engine results conceptual reorganization that uses formal concept analysis. This is done by using redirections to solve conceptual redundancies and by adding preliminary disambiguation and expanding the concept lattice with extra navigation nodes based on Wikipedia's ontology and strong conceptual links.**

## I. Introduction

The Web holds an immense amount of information available in electronic format. Conceptual data analysis is now more helpful than ever for organizing, finding and retrieving the most relevant piece of information in a user-friendly fashion.

Formal concept analysis[1] methods provide ways of structuring search results by identifying clusters of web pages with similar attributes that most likely form a certain concept, a node in a concept lattice described by some subset of the set of attributes or keywords. The best known implementation of such an algorithm used for search result structuring and conceptual navigation is CREDO[2] (Conceptual Reorganization of DOcuments) developed at Fondazione Ugo Bordoni by Claudio Carpineto and Gianni Romano[3]. It runs by querying Yahoo! and organizing the obtained search results into concepts obtained by clustering on the basis of similar attributes.

For example, for the keyword "ruby" CREDO returns the following concept list [2]:



- programming (28)
  - language (14)
  - guide (5)
  - pragmatic (4)
  - rails (3)
  - central (2)
  - tutorials (2)
  - *other (4)*
- language (24)
- rails (16)
- code (14)
- java (10)
- development (8)
- guide (7)
- reilly (5)
- project (5)
- community (5)
- software (5)
- jruby (4)
- application (4)
- tutorial (4)
- book (4)
- *other (22)*

Figure1. CREDO Results for "ruby"

It all seems to work very well, but taking a closer look at these results we notice that all the returned concepts have to do with the Ruby programming language in some way, the problem being that the term "ruby" is not only used to refer to a programming language but it could refer to the gemstone, a person or a location, a song, a game, even a pistol and so on. The fact that the term "ruby" refers to more concepts than just the programming language is involuntarily neglected by the nature of the ordering the search engine provides, on the number of page hits and the first 100 returned results will probably all refer to the programming language known to only a few computer programmers and not to the other meanings of the term familiar to the remaining 99% of the English speaking web surfers.

## II. Formal Concept Analysis Basics and Term-Concept Mappings

Formal Concept Analysis (FCA) is a mathematical theory modeling the concept of "concepts" in terms of lattice theory. To allow a mathematical description of extensions and intensions, FCA starts with a formal context.

A formal context is a triple $K := (G;M; I)$, where G is a set whose elements are called objects, M is a set whose elements are called attributes, and I is a binary relation between G and M (e.g. $I \subseteq X$ ) . $(g, m) \in I$ is read "object g has attribute m" [4].

For $A \subseteq G$, let

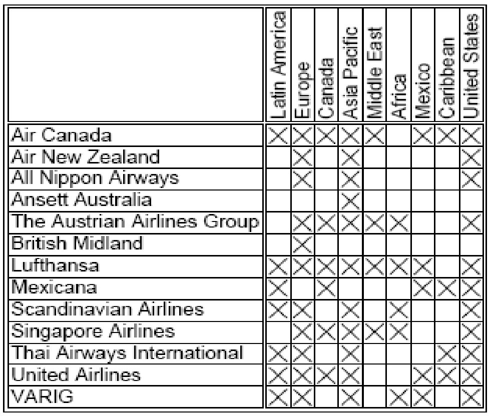
$$A' := \{m \in M \mid \forall g \in A: (g, m) \in I\}$$

and, for $B \subseteq M$ let

$$B' := \{g \in G \mid \forall m \in B: (g, m) \in I\}$$

A formal concept of a formal context (G, M, I) is a pair (A, B) with $A \subseteq G$, $B \subseteq M$, $A' = B$ and $B' = A$. The sets A and B are called the *extent* and the *intent* of the formal concept (A, B), respectively. The s*ubconcept superconcept relation* is formalized by

$(A1, B1) \leq (A2, B2): \Leftrightarrow A1 \subseteq A2 \ (\Leftrightarrow B1 \supseteq B2)$ [4].



| | Latin America | Europe | Canada | Asia Pacific | Middle East | Africa | Mexico | Caribbean | United States |
|---|---|---|---|---|---|---|---|---|---|
| Air Canada | X | X | X | X | X | | X | X | X |
| Air New Zealand | | X | | X | | | | | X |
| All Nippon Airways | | X | | X | | | | | X |
| Ansett Australia | | | | X | | | | | |
| The Austrian Airlines Group | | X | X | X | X | X | | | X |
| British Midland | | X | | | | | | | |
| Lufthansa | X | X | X | X | X | X | X | | X |
| Mexicana | X | | X | | | | X | X | X |
| Scandinavian Airlines | X | X | | X | | X | | | X |
| Singapore Airlines | | X | X | X | X | X | | | X |
| Thai Airways International | X | X | | X | | | | X | X |
| United Airlines | X | X | X | X | | | X | X | X |
| VARIG | X | X | | X | | X | X | | X |

Figure 2. Formal context of airlines and flights in geographic regions [4]

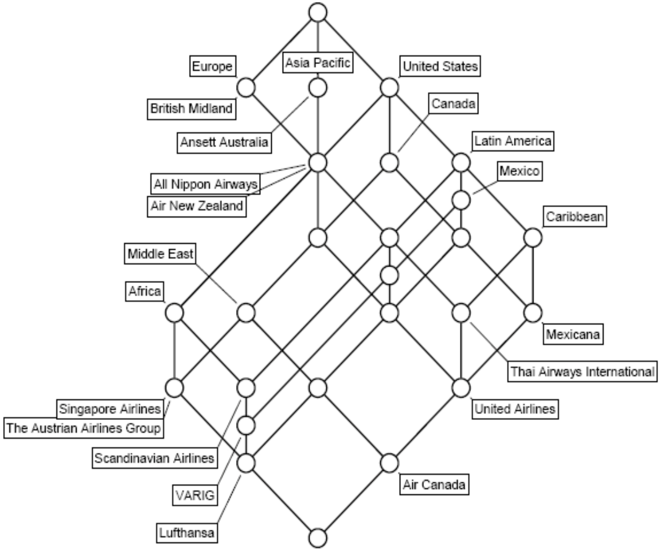


Figure 3. Concept lattice of the formal context in Fig. 2 [4]

The set of all formal concepts of a context K together with the order relation ≤ is always a complete lattice (i.e. for each subset of concepts, there is always a unique greatest common subconcept and a unique least common superconcept), called the *concept lattice* of K, also called *conceptual hierarchy.*

In a line diagram each node represents a formal concept. A concept $c_1$ is a subconcept of $c_2$ if and only if there is a path of descending edges from the node representing c2 to the node representing c1. The name of an object g is always attached to the node representing the smallest concept with g in its extent; dually, the name of an attribute m is always attached to the node representing the largest concept with m in its intent [4].

A term-concept (TC) map is a graph consisting of two types of nodes (terms and concepts), and directed edges that represent relationships between nodes. The TC map represents a bridge from the natural language domain to the concept domain [5].

A more mathematically sound description of these term-concept maps could be given in terms of conceptual graph theory. Multiple terms referring to the same concept (what we shall later discuss as redirection) can be seen as a coreference link (i.e. two or more concepts in a conceptual graph may refer to the same individual). The referents' literals differ but there exists a coreference link between each pair in this group of referents.

One term that links to multiple concepts (this will be discussed under disambiguation) can be seen as the situation in which two or more referents of different types happen to have the same literal. These referents are differentiated based on the type they belong to [6].

## III. WIKIPEDIA REDIRECTION AND DISSAMBIGUATION

Whenever a formal concept analysis paper mentions using an existing ontology to improve the results yielded by clustering objects by attributes and forming concept lattices that can be navigated, the authors most likely mention WordNet as a good source for an ontology. But WordNet doesn't contain proper nouns, which could also refer to concepts of interest to the users of a conceptual search result reorganization application such as CREDO. To include proper terms one would have to make use of an encyclopedia. Fortunately, there already exists an enormous online repository of encyclopedic information: Wikipedia.

Wikipedia also contains about 2 million articles in English which is roughly 20 times the size of WordNet.

Mining the necessary information from Wikipedia is not an easy task as one has to analyze the conventions used to identify different types of pages and internal mechanisms.

Reference [5] provides a good systematization of the structure and mechanisms of Wikipedia required for a term-concept mapping.

Andrew Gregorowicz and Mark A. Kramer classify the pages of Wikipedia in three possible categories:

1. Article
2. Redirect
3. Disambiguation

### *A. Redirection and Concept Renaming*

We will first describe ways of using redirection pages as means of obtaining extra results and properly naming concepts obtained as nodes of a conceptual lattice. Redirection is used to associate multiple terms (or form variations) to a single concept which in this case would be an article page.

To illustrate this mechanism, the queries “john kennedy”, “president kennedy”, “jack kennedy” and “j.f.k.” all redirect to the John F. Kennedy article.

This can be used to name concepts using the most common or most distinguishable form variation or term group. Not all term associations can be transformed into the most common name for that concept, but surely the ones which yielded the most results will have a more common name.

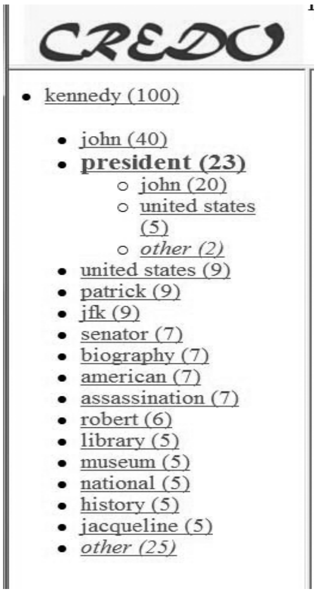


Figure 4. CREDO results for “kennedy” keyword

Considering the “kennedy” example, redirection could be used to merge the CREDO results for “president kennedy”, “john kennedy” and “j.f.k.” under one single conceptual node named “john f. kennedy. As we have explained above, all these terms refer to the same concept but they are not themselves stand-alone concepts. The problem raised by wanting do operate such a renaming is that the concept lattice behind the tree-like menu in CREDO is no longer obvious. Still for purpose of improving conceptual navigation, this mapping of several terms to a single concept through merging nodes (with attributes that redirect to the same concept/article) in the concept lattice or just in the lattice-generated conceptual navigation menu is definitely more intuitive, more human-readable.

The formal description of redirection is the process of clarifying and reducing the formal context of our query. A formal context (G, M, I) is called **clarified**, if for any objects $g, h \in G$, from $g'=h'$ it always follows that $g=h$, and correspondingly, $m'=n'$ for all $m,n \in N$. Even if a page named J.F.K. actually does not independently exist, by redirection we might assume that the set of key words defining it is the same as for John F. Kennedy. **Reducing** the context is the removing of **reducible attributes**, i.e., of objects with **inf**-reducible attribute concepts and of **reducible objects**, i.e., of objects with **sup**-reducible object concepts.

### *B. Disambiguation and Expanding the Concept Lattice*

Disambiguation Wikipedia pages provide a different kind of term-concept resolution. Disambiguation is needed when a term refers to more than one concept. In the Introduction of this article we gave an example of CREDO results when querying with the keyword “ruby”. We noticed the problem with these results was that by considering only the first 100 results Yahoo! provided, the lattice-generated conceptual navigation menu only referred to the Ruby programming language and not any other meaning like the most common one, that of a red gemstone.

Wikipedia disambiguation pages offer alternatives for terms/keywords that can point to multiple concepts/article pages.

TABLE 1
Disambiguation for “ruby” in Wikipedia

A **ruby** is a red gemstone
**Other uses**
Ruby (programming language), a computer programming language
Ruby on Rails, a Ruby based web development framework
Ruby characters, a way to show the pronunciation of logographic characters
Ruby (annotation markup), the implementation of Ruby characters in XHTML
Ruby (hardware description language), a language for designing computer circuits
Rubies of Eventide, an MMORPG
Ruby laser
Ruby (elephant), an animal at the Phoenix Zoo famous for her paintings
Ruby pistol, a French WWI sidearm manufactured exclusively in Spain
Ruby Murray, Cockney rhyming slang for curry, from the singer Ruby Murray
Rubies (ballet), the second movement of George Balanchine's Jewels

**People**
Ruby (Egyptian singer) (born 1981)
Ruby Dee (born 1924), actress
Ruby Lin (born 1976), Taiwanese actress
Ruby Murray (1935–1996), singer
Ruby Walsh (born 1979), Irish jockey
Ruby Wax (born 1953), comedian
Jack Ruby (1911–1967), the man who killed Lee Harvey Oswald
Lloyd Ruby (born 1928), race car driver
Ruby Dandridge (born in 1899), Actress
Sam Ruby, Software developer

**Locations**
Ruby, Alaska
Ruby, Arizona
Ruby, New York
Ruby Dome, the highest peak of the Ruby Mountains
Ruby Mountain, small stratovolcano in Stikine Region, British Columbia, Canada
Ruby Mountains, mountain ranges in the western United States

**Entertainment**
Ruby (Supernatural), a character in the third season of the series Supernatural
Ruby (band), a Scottish electronic band
Ruby (film), a 1992 film about Jack Ruby starring Danny Aiello and Sherilyn Fenn
Ruby (1977 film), a horror film by Curtis Harrington and starring Piper Laurie
Ruby (talk show), a British television chat show hosted by Ruby Wax
Ruby the Galactic Gumshoe, the titular character of a radio show
Ruby Gloom, a stationery franchise and animated TV series
Ruby, a fictional character from The Tribe played by Fleur Saville
Pokémon Ruby, a video game
Ruby (Pokémon), a main character in the manga Pokémon Adventures
Ruby (character), a main character in According to Jim played by Taylor Atelian
"Ruby, Don't Take Your Love to Town", a 1969 hit country song by Kenny Rogers a
Ruby Trollman, a character from the animated series Trollz
"Ruby Tuesday" (song), a song by The Rolling Stones
Ruby (The Land Before Time), a character on The Land Before Time TV series
Ruby Records, a record label
"Ruby" (song), a 2007 song by Kaiser Chiefs
Ruby, a fictional character used by ATI for advertising video cards
"Ruby" (song), a 1953 song written by Richard Hayman for the movie Ruby Gentry
"Through The Eyes of Ruby", a song by the Smashing Pumpkins.
Ruby(cartoon) A fictonal character from the cartoon, Max & Ruby.
Winter Wonderland Ruby Version, a Months film game.

**Literature**
Ruby (novel), an 1889 novel by Amye Reade
Ruby (Andrews novel), a 1994 novel by V.C. Andrews

Surely the variety and the number of concepts the term “ruby” can point to is impressive, especially when one thinks of the results CREDO conceptual reorganization offered. It follows that the formal context and implicitly the concept lattice need to be expanded by adding objects (web pages) obtained by supplementary queries aimed at collecting extra concepts by adding disambiguation keywords for each case in new searches. These keywords can easily be extracted from a disambiguation page like the one in Table 1. Although such an approach can be costly with regard to time consumption, it would surely bring conceptual navigation closer to what naïve users think it should be. An exploitable advantage is the fact that Wikipedia provides taxonomic information by encapsulating results into different categories specific to its structure[7] but also very familiar to any web surfer (types in terms of conceptual graphs theory: People, Locations, Entertainment, Literature etc.).

The conceptual navigation system generated by such an extension would be certainly more user-friendly, more human-readable, but, again, slightly further away from the algebraic structure behind CREDO-like search results reorganization.

In fact, disambiguation provides a multilevel navigation structure, very close to the nested line diagram navigation concept, provided by Formal Concept Analysis. Moreover, the nested line diagrams could be a useful tool to visualize disambiguation categories (Fig. 5).

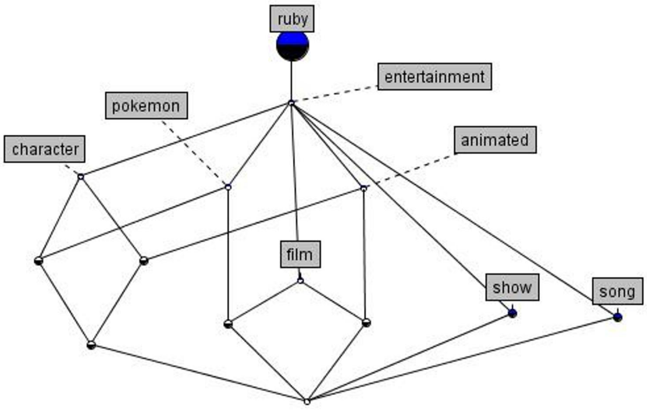


Fig.5 Conceptual hierarchy for entertainment disambiguation category

## III. STRONG LINKS AND RELATED CONCEPTS

The Wikipedia article page can be assimilated to the concept when talking about term-concept mappings. As we have already seen, multiple terms can point to the same concept (in which case we can use renaming by means of Wikipedia redirection pages) or a single term can point to multiple concepts (in which case we can use Wikipedia disambiguation pages to provide all the existing concepts referred by that term). Still, Wikipedia can provide one more valuable type of relation, a relation between two concepts. This however is rather difficult to mine.

All Wikipedia articles contain links to other relevant articles. These links can be grouped in two distinct categories [5]:

1. Unidirectional links
2. Bidirectional links

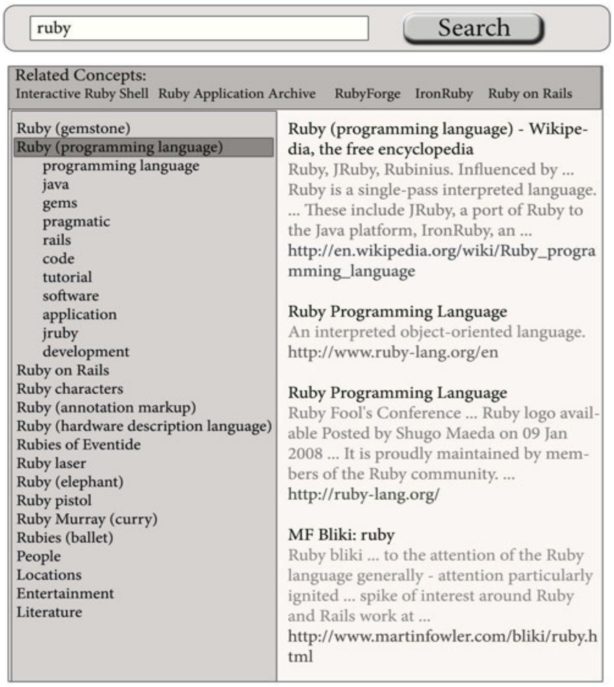
The unidirectional links (i.e. links to other articles that do not link back to the current article) usually define an inclusion relation between the two concepts or a relation of little relevance. They are weak links.

The bidirectional links (i.e. links to other articles that in turn link to the current article) however define strong relations between concepts. This is why, they could be considered useful when trying to develop a conceptual navigation system. Still, it is almost impossible to define such relations between concepts in an already specified formal context.

The best integration of this extra information into a CREDO-like system would be a secondary menu that allows navigation from already renamed concepts to related searches.

## IV. IMPROVED CONCEPTUAL NAVIGATION SYSTEM

As a conclusion to the term-concept mapping features used in improving conceptual Web navigation systems presented above, we now describe a CREDO-like real-time conceptual navigation system. The primary navigation and means of refinement is a tree control derived from the topmost levels of the iceberg concept lattice built from the terms extracted from the search results.



Related Concepts:
Interactive Ruby Shell Ruby Application Archive RubyForge IronRuby Ruby on Rails

Ruby (gemstone)
Ruby (programming language)
programming language
java
gems
pragmatic
rails
code
tutorial
software
application
jruby
development
Ruby on Rails
Ruby characters
Ruby (annotation markup)
Ruby (hardware description language)
Rubies of Eventide
Ruby laser
Ruby (elephant)
Ruby pistol
Ruby Murray (curry)
Rubies (ballet)
People
Locations
Entertainment
Literature

Ruby (programming language) - Wikipedia, the free encyclopedia
Ruby, JRuby, Rubinius. Influenced by ... Ruby is a single-pass interpreted language. ... These include JRuby, a port of Ruby to the Java platform, IronRuby, an ...
http://en.wikipedia.org/wiki/Ruby_programming_language

Ruby Programming Language
An interpreted object-oriented language.
http://www.ruby-lang.org/en

Ruby Programming Language
Ruby Fool's Conference ... Ruby logo available Posted by Shugo Maeda on 09 Jan 2008 ... It is proudly maintained by members of the Ruby community. ...
http://ruby-lang.org/

MF Bliki: ruby
Ruby bliki ... to the attention of the Ruby language generally - attention particularly ignited ... spike of interest around Ruby and Rails work at ...
http://www.martinfowler.com/bliki/ruby.html

Figure 6. An improved result page for conceptual navigation on “ruby” keyword containing the described features

Features:

1. Concept renaming (merges two or more nodes into one real concept name given by Wikipedia redirection).
2. Expanding the formal context with concepts referred to by the keyword (e.g. Ruby on Rails, Ruby characters, Rubies of Eventide, Ruby laser, Ruby pistol, Ruby-list of people, Ruby- list of locations etc.).
3. A secondary navigation menu consisting of related concepts detected by exploring the strong links of the current concept (by Wikipedia redirection).

As it can be seen, the search result reorganizer described above has a wider range of conceptual results associated with the search keyword. This makes search result navigation and refinement easier. Also, having in view the continuous growth of Wikipedia, our application's performance will automatically improve, yielding new term-concept mappings in its navigation system as they appear in our culture. Although mining Wikipedia is not a new idea, mining the more subtle inner structure (the term concept mappings) and using it to improve FCA – based conceptual search result reorganizer surely is an innovation.

Rather than building on-the-spot conceptual neighborhoods like SearchSleuth[8], our approach preserves the simplicity and performance of CREDO and at the same time provides broad contexts and conceptual neighbors mined from concept-to-concept relations (bidirectional links).

New directions for development could take advantage of Wikipedia's richness by branching searches through Wikipedia search which usually implies disambiguation for broad terms.

The advantage of using Wikipedia rather than some ontology specially designed for use within FCA applications is that its taxonomical structure and inter-concept relations although vague, sometimes inconsistent and more difficult to mine is backed by an ever-growing enormous amount of data. And clearly with this great amount of data comes great reliability.

## References

[1] Formal Concept Analysis Homepage, http://www.upriss.org.uk/fca/fca.html

[2] CREDO, http://credo.fub.it/

[3] C. Carpineto and G. Romano, "Concept Data Analysis: Theory and applications," John Wiley & Sons, September 2004

[4] B. Ganter and G. Stumme "Formal Concept Analysis: Methods and Applications in computer Science," TU Dresden, Summer 2003

[5] A. Gregorowicz and M. A. Kramer "Mining a Large-Scale Term-Concept Network from Wikipedia," MITRE Corporation, 202 Burlington Road, Bedford, USA, August 2006

[6] Online Course in Knowledge Representation using Conceptual Graphs http://www.huminf.aau.dk/cg/

[7] Semantic MediaWiki project. http://en.wikipedia.org/wiki/Semantic_MediaWiki

[8] J. Ducrou and P. Eklund "SearchSleuth:The Conceptual Neighbourhood of an Web Query", Schol of Computer Science and Software Engineering, University of Wollongong