Mining a Large-Scale Term-Concept Network from Wikipedia

Andrew Gregorowicz and Mark A. Kramer

MITRE Corporation, 202 Burlington Road, Bedford, MA 01730 USA
{andrewg@mitre.org, mkramer@mitre.org}

Abstract. Social tagging and information retrieval are challenged by the fact that the same item or idea can be expressed by different terms or words. To counteract the problem of variable terminology, researchers have proposed concept-based information retrieval. To date, however, most concept spaces have been either manually-produced taxonomies or special-purpose ontologies, too small for classifying arbitrary resources. To create a large set of concepts, and to facilitate terms to concept mapping, we introduce mine a network of concepts and terms from Wikipedia. Our algorithm results in a robust, extensible term-concept network for tagging and information retrieval, containing over 2,000,000 concepts with mappings to over 3,000,000 unique terms.

Keywords: Information retrieval, concept search, Wikipedia, text mining.

1 Introduction

There is a rich, many-to-many mapping between terms (names, words, and phrases) and concepts (things and ideas). A single thing or idea can be expressed by synonymous terms or phrases, or known by multiple names, including variations, abbreviations, acronyms, nicknames and aliases. Conversely, the same word, name, or phrase can represent different things, people, or ideas. This is a problem in information retrieval because text-based retrieval fails to differentiate between multiple senses of a term. Ranking search results is not a solution for finding resources related to a specific sense of a word. A similar problem occurs in social tagging, where users select natural language terms to describe items such as photographs and documents. Research has observed that, on average, people have less than 20% chance of choosing the same word to describe the same concept [1].

Consider searching for the term “ruby”. The well-known concepts associated with this term are represented by three definitions in WordNet [2] (the gemstone, the color, and the mineral), and 23 interpretations in Wikipedia (including the programming language, the assassin, the Egyptian singer, the video game, as well as towns, novels, films, and record albums). Searching the tag “ruby” at the photo-sharing site Flickr recalls many concepts: people and pets named Ruby, gems, computer screens showing Ruby code, ruby-colored slippers, landscape shots of Ruby Beach (WA), and more. Refining the search with additional keywords may not entirely solve the

problem: for example, the term “Ruby Falls” recalls Anna Ruby Falls in Georgia, a 1990’s feminist rock quartet from New York, and a 145-foot underground waterfall located within Lookout Mountain, near Chattanooga, TN.

Several authors have suggested concept search as a potential approach to this problem, in which search is performed on a specific sense of a term, for example, the “red corundum” sense of ruby. Analogously, we propose concept tagging, where social tags are drawn from a pre-existing concept space, avoiding the ambiguity associated with terms. Concept search has four essential requirements:

1. A concept space providing topical coverage for a defined set of resources,
2. A mapping of resources into the concept space,
3. An information requirement expressed in terms of the concept space, and
4. A search mechanism to retrieve information based on the target concepts.

Dealing with concepts, rather than terms, improves precision by eliminating non-intended senses of the search terms. In addition, concept search spawns powerful extensions based on relationships between concepts. For example, thesaurus relationships between concepts such as related, broader, and narrower provide the basis for query refinement, which can improve recall and precision [3] [4]. Deeper semantic relationships, à la semantic web, lead to applications such as semantic querying, rule-based inference, and question answering [5] [6].

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. In terms of the four requirements for concept search, the TC map provides an enumeration of concepts, a resource for document classification, and method for the user to locate the concepts relevant to an information requirement. Finding the relevant concepts by direct browsing may not be feasible, especially since the names chosen to represent concepts may not be intuitive to a user (should the concept for ruby-the-rock be named “RubyGemstone” or “red_corundum”?) Rather than relying on naming conventions, we envision the user entering natural-language text, and the system responding with a list of candidate concepts. The generation of the candidate concept list cannot be based on alphabetic look-ahead or other text considerations. For example, the term “ruby” should evoke the candidate concept “RedCorundum” despite the lack of co-occurring text. For this purpose, a TC map is required.

In this paper, we address the problem of generating a large, general-purpose TC map. By general purpose, we mean inclusive knowledge of the world, rather than knowledge in a focused domain. In this paper, we limit the scope of our investigations to TC maps composed of common and proper nouns. While this is not a fundamental restriction, it is generally accepted that open grammatical classes, particularly nouns, are better than other parts of speech for indexing resources. Furthermore, we address only the structure of the map, not the determination of its parameters (frequencies of terms or concepts, conditional probabilities of terms given concepts, etc.) Although the concepts here could be applied to other languages, in this paper, we limit our scope to the English language. We also do not address the problem of mapping resources to the concept space.

This paper is arranged as follows: In Section 2, we consider various alternatives sources for generating a TC map, including dictionaries, thesauri, and encyclopedias. In Section 3, we present an analysis of the structure of the largest on-line encyclopedia, Wikipedia, and present in Section 4 a method of mining Wikipedia data into large network of terms and concepts. Section 5 presents results, and Section 6 concludes with next steps and potential improvements.

2 Formulating Term-Concept Maps

2.1 Ontology

There are several ways a TC map could be represented. Obrst [7] explains the choice in terms of an "ontological spectrum" that ranges from weak to strong semantics. Taxonomies, involving only generic parent-child relationships, are at the weak semantics end of the ontology spectrum. At the other end are rich semantic models pertaining to classes of real-world things and their relationships, for example, canCut(scissors, paper). Between these extremes are thesaurus-like models involving lexigraphic relationships such as synonyms, hypernyms, and hyponyms.

Our model includes terms and concepts as two fundamental, disjoint classes. Relationships can connect terms to other terms (TT), concepts to concepts (CC), or terms to concepts (TC). Since the purpose to navigate from terms to concepts and visa-versa, we are primarily focused on modeling TC relationships. We exclude TT relationships, which are entailed by CC and TC concepts. For example, if a concept C1 has two alternative terms T1 and T2, there is an implied synonymy relationship between T1 and T2.

To model TC relationships, we use a subset of SKOS, the Simple Knowledge Organization System [8]. While not sufficient for question answering and other advanced applications, SKOS is sufficient to represent simple TC maps. SKOS provides a single class for concept nodes, *skos:concept*. TC relationships are defined by three mutually-exclusive properties of concepts:

- *skos:prefLabel* – the preferred term for a concept
- *skos:altLabel* – alternative terms for the concept including acronyms, abbreviations, spelling variants, and irregular plural/singular forms.
- *skos:hiddenLabel* – terms that should not appear in the user interface, but may be used in free text search operations; typically used for common misspellings

In SKOS, labels can be language-specific, so an invariant core of concepts can be expressed in terms drawn from multiple languages. For CC relations, SKOS provides thesaurus-like relations, including *skos:broader*, *skos:narrower*, and *skos:related*. While more complex CC relationships can be considered, we do not explore this possibility further in the current paper. In addition, SKOS provides *skos:isPrimarySubjectOf* and *skos:isSubjectOf* to tag resources with concepts. While SKOS is designed as an RDF dialect, the limited set of relations can be represented in a traditional database, which we do here because of the large size of our TC map.

2.2 Using Wikipedia as the Source of Terms and Concepts

One can conceive of various ways to create term-concept maps. For example, we could begin with a manually-created taxonomy, representing the concepts, and attach terms to each node of the taxonomy. Taxonomies are useful in specialized domains where there is a coherent community of users able to agree upon the organization of concepts, however, for the representing real world items and their relationships, there is no adequate taxonomy or universal organizational scheme with enough depth for our purposes.

Another possibility is to leverage dictionaries and thesauri such as WordNet [2], a lexical database containing English-language words, definitions, and relations. WordNet arranges terms with like meaning in groups called synonym sets (synsets), which are analogous to concepts. Synsets are related to other synsets by semantic relations, including hypernym (broader), hyponym (narrower), holonym (contains), and meronym (part of). WordNet 2.1 contains 117,097 unique noun terms, arranged into 81,426 synsets. While this covers most English common nouns, there is poor coverage of proper nouns. Since most tags, searches, and categorizations relate to proper nouns rather than common nouns, WordNet is not an ideal source for a general-knowledge TC map.

Since proper nouns form the majority of tags, it follows that an encyclopedia is the proper source for the concept space. Wikipedia is unique in that (1) it provides a broad source of general knowledge, (2) it has considerable depth in many specialized fields, and (3) its development model allows it to keep up with a rapidly-changing world. Each article in Wikipedia corresponds to the current notion of a concept. If you enter a verb, for example “run”, it evokes the gerund noun form “running” and other noun forms such as a baseball run, bank run, etc. At 2.1 million, the number of concepts in Wikipedia is more than 25 times larger than WordNet. Both common nouns and proper nouns are included, but given that the number of English common nouns is only about 100,000, it is clear that the vast majority of Wikipedia articles relate to proper nouns.

Concepts are only half the story, however, since our goal is to associate terms with concepts. What constitutes a term in Wikipedia, and the relation of those terms to concepts, is less obvious. When you type in a term, you may be directed to an article whose title is different than the search term. For example, the terms “John Kennedy”, “Jack Kennedy”, and “President Kennedy” all redirect to the article on “John F. Kennedy”. In other cases, a term may invoke a disambiguation page, showing that the term is related to multiple concepts. Clearly, Wikipedia implicitly contains an extensive TC map; the question is how to mine the data.

3 Wikipedia Structure

The Wikipedia is built on the MediaWiki software package [9]. The MediaWiki package provides various XML-formatted data extractions, which are made available by Wikipedia [10]. The extraction we chose to work with contains only current article

content, image descriptions, templates and primary (non-meta) pages. Extractions that provide additional data, such as page history, are also available.

The core of the MediaWiki XML format is the page element. The following illustrates the basic structure:

```
<page>
        <title>Title of Page</title>
        <revision>
                {several tags omitted for brevity}
                <text> A bunch of text here </text>
        </revision>
</page>
```

Pages fall into one of three possible categories:

1. **Article**: user-visible text that describes a single concept
2. **Redirect**: a non-visible pointer to another Wikipedia page
3. **Disambiguation**: a visible page enumerating concepts described by a term

A weakness of MediaWiki from the perspective of data mining is that many page properties, including page categories, are not indicated by XML tags; rather, they must be deduced based on a combination of naming conventions, embedded keywords, and special symbols. For example, a redirect page is a page with no text other than a directive in the form #REDIRECT link. The most common purpose of a redirect page is to associate a search term to an article. For example, a redirect page is used to associate the term "Jack Kennedy" to the article entitled "John F. Kennedy". Links have a special syntax, easily parsed out of the page text, denoted by double brackets, [[page name | link term]], where page name is the destination of the link, and the optional link term is the visible text. In reference to TC maps, titles of redirect pages represent alternate terms (*skos:altLabel*) for the concept represented by the destination.

A disambiguation page describes multiple concepts associated with a term. Most (but not all) disambiguation pages are signaled by the string "(disambiguation)" in the page title, or the category may be signaled by the string "{{disambig}}" in the text. Wikipedia uses two common navigations for disambiguation. The first deposits the user directly on a disambiguation page, and is used when none of the interpretations of the term are much more common than the others. The second navigation style takes the user to a default article, representing the most common interpretation of the term, which contains a link to disambiguation page for other uses of the term. In both cases, the disambiguation page, minus the "(disambiguation)" extension, represents an alternate term (*skos:altLabel*) for each concept enumerated on the page. The title default page, if it exists, represents the preferred term (*skos:prefLabel*) for the article found there.

The remaining pages are articles. Each article represents a unique concept, and the title is taken as the preferred term (*skos:prefLabel*) for that concept. The concept name is taken as the same as the title of the article.

In addition to the features discussed here, Wikipedia pages sport many features that could potentially enrich CC and TC relationships, including table of contents, subheadings, data tables, and external references. Exploitation of additional structure remains for future work.

4 Algorithm

The creation of the TC map employs a three-pass algorithm:

First Pass: This pass creates all terms and concepts. All page titles are taken to be terms, and inserted into the database. If a term contains the text string "(disambiguation)" it is stripped from the term. Conventions like connecting underscores and omitted spaces between words are replaced by whitespaces. For each article, a concept is created with the same name as the term, and a *skos:prefLabel* link is established between the term and concept. All other page content is ignored.
Second Pass: Only redirection pages are considered. By examining the page target of the link (following redirect chains if necessary), the concept associated with the link term can be determined, and therefore the term from the page title can be linked to the concept represented by the redirect target term. This pass creates links between terms and concepts, establishing synonyms (*skos:altLabel*) for terms representing a concept.
Third Pass: Only disambiguation and articles (concept pages) are considered. For disambiguation pages, all links are followed to the underlying concept, which establishes disambiguation page titles as homonyms (*skos:altLabel*). In addition, concept pages are examined for all links. Again, targets for the links are used to determined linked concepts, allowing for the creation of concept-to-concept links (*skos:related*).

An application was written to implement the parsing algorithm and populate a relational database logically equivalent to SKOS, which could then be queried to explore the TC map. The Wikipedia extraction examined in this paper (taken May 2006) was obtained as a single XML file, compressed using the bzip2 data compressor [11]. The compressed file is approximately 1.4 GB in size which expanded to 5.4 GB.

The application was written in Java, which was compiled and run on Sun's Java SE 5 Platform. Extraction of the information from the Wikimedia XML was performed via the SAX API [12] using Apache Xerces 2.8.0 [13] as the underlying parsing engine. PostgreSQL 8.1.3 [14] was used as the relational database to store the TC map. Hibernate 3.1.3 [15], an Object-Relational Mapping tool, was used to reduce the amount of code needed to interface with the database.

As shown in Figure 1, the TC map is stored in four tables in the relational database: *terms*, *concepts*, *term_concept_relationships* and *concept_relationships*. The *terms* table stores the terms extracted from Wikipedia. The *terms* table has a numeric identifier for each term, the term itself and a boolean indicating if the term is a preferred term for a concept. The *concepts* table is very similar to the *terms* table with a numeric identifier for the concept and the concept name.

The *term_concept_relationships* table stores the mappings between terms and concepts in terms of the numerical identifiers. It then contains the id of the term and the id of the concept that are related. Foreign key constraints are enforced to ensure that a valid id is provided for both the concepts and terms.

The *concept_relationships* table maintains the mappings between concepts. This table contains a unique numeric identifier for the relationship, the id of the concept that makes the reference and the id of the concept that it refers to. This relationship is created when the text of a Wikipedia article links to another Wikipedia article. In this case, the concept described by the article text is considered the "from concept" while the concept that is the target of the link is the "to concept". If the relationship is bidirectional, which means that both articles refer to each other in their texts, two records will be created in the *concept_relationships* table with the "to" and "from" concept ids transposed. The only CC relationship used at this time is *skos:related*.

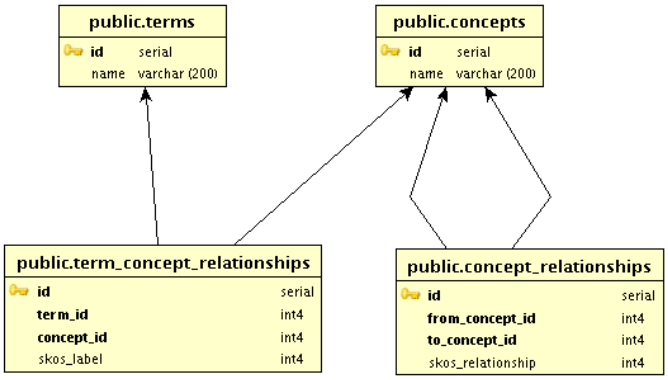

Fig. 1. Database structure corresponding to SKOS.

5 Results

To test the results of the algorithm, we confirmed by manual spot-checking that the correct terms and concepts network had captured the implicit term-concept network in Wikipedia. Initially, we found spurious TC links, which were traced down to the appearance of additional links on disambiguation pages, not indicating related concepts. An example of spurious disambiguation links is shown in the snippet from the Java disambiguation page, below:

***Java** may also mean:*

- *Java (band)*, a French band
- *Java (board game)*, a board game set on the island of Java
- *Java (cachaça)*, a brand of cachaça
- *Java (dance)*, a Parisian *Bal-musette* dance
- *Java Man*, one of the first specimens of Homo Erectus to be discovered
- *"Java" (song)*, a song by *Allen Toussaint*. Al Hirt also recorded a popular version, which was later used on The Muppet Show

The links to Bal-musette and Allen Toussaint are not concepts directly associated with the term "java". This problem was eliminated by filtering outgoing links to include only those links directly following bullets, leveraging a widespread Wikipedia convention.

We then examined the CC relationships, and observed that many CC relationships extracted from Wikipedia links represented (subjectively) weak relationships. We hypothesized that strong relationships would be signaled by bidirectional links. For example, since the word "ruby" is derived from Latin, there is a link to "Latin" in that article, but the article on "Latin" has no link to ruby. Since *skos:related* is a symmetric property, we reasoned that Wikipedia links in both directions were needed to instantiate a *skos:related* relationship. Intuitively, this is borne out for Ruby:

- **Bidirectional Links (strong):** Gemstone, Corundum, Aluminum Oxide, Chromium, Verneuil Process, Sapphire, Rutile, Asterism, Rajaranta Ruby, Neelanjali Ruby, Laser, Ephod, Birthstone, July
- **Monodirectional Links (weak):** Mineral, Latin, Africa, Asia, Australia, Greenland, Myanmar, Sri Lanka, Kenya, Madagascar, Thailand, U.S. State, Montana, N. Carolina, S. Carolina, Mogok Valley, Spinel, Hardness, Mohs Scale, Diamond, Dopant, Gemology, Cabochon, Carat, Bangalore, India, Switzerland, Rebbenu, Bachya, Exodus, Reuben, Hebrew, Proverbs

Therefore we decided to re-run the algorithm, including creating a *skos:related* link only when the link was bidirectional. A depiction of a small part of the resulting network, starting with the term "Java", at different scales, is shown in Figures 2 and 3.

Fig. 2. Concepts directly related to the term "java" derived from Wikipedia

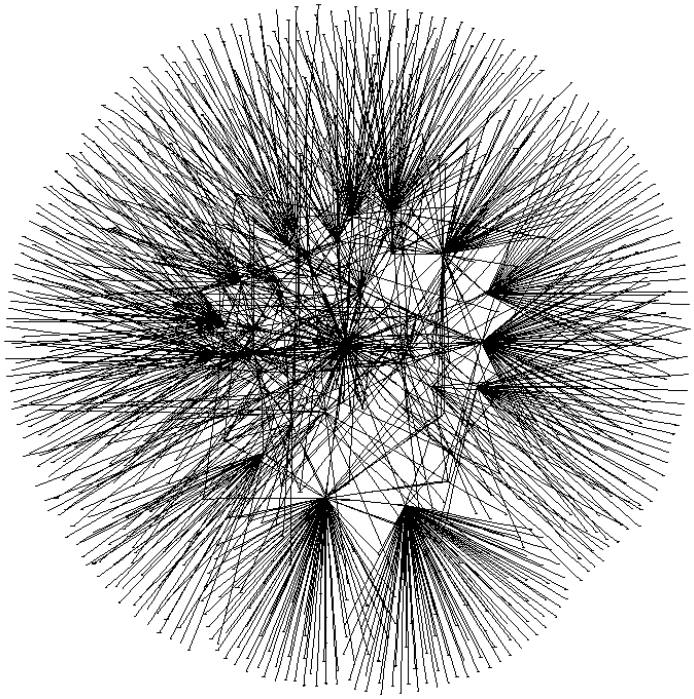

Fig. 3. Term-concept map from Wikipedia, with central point "java"

6 Conclusions and Next Steps

In this paper, we have shown how a large network of terms and concepts can be extracted from Wikipedia. The resulting concept space can be used as a source of unambiguous tags to index resources. Represented as RDF using SKOS representation, the network can be searched using semantic query languages such as SPARQL, or as a conventional database, through SQL.

Wikipedia is a rich source of semi-structured information, and techniques to mine this resource should continue to be developed. The current study only scratches the surface of what is possible via "Wikipedia Mining", particularly in the area of concept-concept relationships. Because Wikipedia is loosely-structured, this pursuit relies heavily on conventions within the Wikipedia community, and requires introduction of heuristics to yield good results. The thrust of several semantic wiki projects is to make information more accessible through use of additional semantic tags [16].

Information from Wikipedia could be augmented with information from other sources. Other language-specific versions of Wikipedia could be included in the same term-concept map, to form a multi-lingual resource. Dictionary and thesaurus

information could be included to capture common nouns and other parts of speech, broader and narrower relationships, additional acronyms from sites like Acronyms.com, and more. By applying natural language understanding to the article contents, rich ontological relationships between concepts could be developed. In addition, we have not explored the statistical quantification of relationships that could help rank the most likely concepts given one or more terms.

References

1. Furnas, G.W., et al. The Vocabulary Problem in Human-System Communication, Communications of the ACM, 30 (1987) 964-971
2. Fellbaum, C.: *WordNet: An Electronic Lexical Database*. MIT Press (1998)
3. Tudhope, D., C. Binding, D. Blocks, D. Cunliffe, FACET: Thesaurus Retrieval with Semantic Term Expansion, Joint Conference on Digital Libraries (2002)
4. Biddulph, M., A Semantic Web Shoebox – Annotating Photos with RSS and RDF, Proc. World Wide Web Conference (2003)
5. Ben Necib, C. and J. Freytag,Semantic Query Transformation using Ontologies Proc. Intl. Database Engineering & Application Symposium (2005)
6. Hess, C., and M. DeVries, From Models to Data: a Prototype Query Translator for the Cadastral Domain, Joint Workshop on Standardization in the Cadastral Domain (2004)
7. Obrst, L. Ontologies for semantically interoperable systems, ACM International Conference on Information and Knowledge Management (2003) 366-369
8. SKOS, http://www.w3.org/2004/02/skos/
9. MediaWiki, http://meta.wikimedia.org/wiki/MediaWiki
10. Wikimedia database dump service. http://download.wikimedia.org/enwiki/
11. Bzip2 data compression. http://www.bzip.org/
12. SAX: Simple API for XML. http://www.saxproject.org/
13. Xerces XML Parser. http://xerces.apache.org/
14. Postgres Open Source Database. http://www.postgresql.org/
15. Relational Persistance for Java and .NET. http://www.hibernate.org/
16. Semantic MediaWiki project. http://en.wikipedia.org/wiki/Semantic_MediaWiki