WikiPics: Multilingual Image Search based on Wiki-Mining

Daniel Kinzler
Wikimedia Deutschland
Eisenacherstr. 2
10777 Berlin, Germany
daniel.kinzler@wikimedia.de

ABSTRACT
This demonstration introduces WikiPics, a language-independent image search engine for Wikimedia Commons. Based on the multilingual thesaurus provided by WikiWord, WikiPics allows users to search and navigate Wikimedia Commons in their preferred language, even though images on Commons are annotated in English nearly exclusively.

Categories and Subject Descriptors
H.3.1 [Information storage and retrieval]: Content Analysis and Indexing – *Dictionaries, Thesauruses.* H.3.3 [Information Storage and Retrieval]: Information Retrieval – *Clustering, Search process, Selection process.* H.3.5 [On-Line Information services]: *Web-based services*.

Keywords
Wikipedia, Search, Thesaurus, Translation, Dictionary, Multilingual, Navigation.

1. INTRODUCTION
WikiPics aims to make the over 6 million images on Wikimedia Commons accessible in the user's native language. It serves as a demonstration of how information mined from a massive, collaboratively maintained lexical semantic resource like Wikipedia can be applied to overcome the language barrier.

Images on Commons are described largely in English, and the categorization scheme is exclusively English. To overcome this problem, a multilingual dictionary or thesaurus is needed.

2. LOOK-UP
A multilingual thesaurus is a powerful tool to map words from different languages to abstract, language-independent meanings. Wikipedia has been found to be particularly well suited as a basis for automatically construction such a thesaurus [1][4][6][7]. WikiWord builds such a thesaurus from Wikipedias in different languages [2]. In the process, it also associates wiki pages and categories with each concept entry. This way, it provides the basis for searching and navigating each of the wikis in any language – specifically, in the case of WikiPics, the navigation of Wikimedia Commons.

WikiSym '10, July 7-9, 2010, Gdańsk, Poland

When the user enters a term in WikiPics, along with the information which language that term is in, WikiPics performs a look-up on the WikiWord thesaurus via a web API. The result of this look-up is a list of concept entries, each representing one meaning of the term, along with information about how often that term is used in Wikipedia to refer to this concept (*signification frequency*). For later use, the concept entry also contains a short definition (*gloss*) of the concept (taken from the first sentence of the corresponding Wikipedia article in the given language), as well as relations to other concepts. These relations, such as broader, narrower, similar or related concepts, may later be used for navigation.

The concept entry returned by WikiWord also contains references to the Wikipedia pages that cover this concept in different languages. Note that Wikimedia Commons is treated like a Wikipedia here, with "commons" being handled as a pseudo-language.

Thus, WikiPics knows the pages and categories on Commons that correspond to each possible meaning of a given term. It can then easily collect the images on these pages and categories, and display them to the user: one list of images for each possible meaning of the search term.

3. RANKING
Ranking of images is governed by two factors: where the image is used, and how the image itself is rated. Image usage on Wikipedia articles is considered to be particularly relevant, following the assumption that these pages will contain a small selection of particularly helpful images. Thus, images used on the Wikipedia articles that correspond to a given topic (concept) will be ranked higher.

However, some images are used on a great many Wikipedia pages. These tend to be icons and indicators that are not related to the page's content, and are thus likely to be irrelevant to the search. Following this assumption, images used on more than some fixed number (e.g. 10) of pages on a single Wikipedia are ignored.

The other criterion, image rating, is applied by a variety of community processes for accessing image quality. Among other things, Commons tags images as "featured", "valuable", or "quality". Images with quality indicators like this will be ranked higher by WikiPics. Contrarily, tags indicating problems like low resolution, etc., may cause an image to be ranked lower.

4. RELATED WORK
As mentioned, Wikipedia has been found to be a valuable resource for thesaurus construction. Some attempts have been made to use such a thesaurus to power a search engine for Wikipedia, most notably Koru [3], which features a rich user interface for navigation, and the more basic WikipediaThesaurusV2 [4]. Both provide a similar mode of

navigation to related topics like WikiPics does, but neither provides a mode for searching images. Also, they do not appear to combine information from multiple languages, but rather use a separate thesaurus for each language.

One approach that does transfer information between languages is CL-ESA [7], which is however not based on a thesaurus, but rather a traditional retrieval system using document similarity.

The search engine Wikiwix by the French company Linternet features a mode for searching Wikimedia Commons, and it does seem to use the Wikipedia pages in the user's language to do so. It does not, however, provide a choice between different meanings of a term, but provides a traditional "flat" list of search results. Since no publications are available about how Wikiwix works, a detailed comparison to WikiPics is not possible.

5. DEMONSTATION

The demonstration will cover an overview of WikiPics' capabilities, such as search and navigation, as well as examples illustrating the way images are ranked. It will be conducted as an on-screen demonstration of the live web site, with an extended question-and-answer part. The demonstration is expected to take 60 to 90 minutes.

6. REFERENCES

[1] Gregorowicz and Kramer 2006. *Mining a Large-Scale Term-Concept Network from Wikipedia*. Technical Report, Mitre.

[2] Kinzler 2008. *Automatischer Aufbau eines multilingualen Thesaurus durch Extraktion semantischer und lexikalischer Relationen aus der Wikipedia*. Diploma Thesis, ASV, University of Leipzig.

[3] Milne and Nichols 2007. A Knowledge-Based Search Engine Powered by Wikipedia. *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, ACM.

[4] Nakayama et al. 2007. Wikipedia Mining for an Association Web Thesaurus Construction. *WISE* vol. 4831, Springer.

[5] Nakayama et al. 2008. A Search Engine for Browsing the Wikipedia Thesaurus. *Database Systems for Advanced Applications*, Lecture Notes in Computer Science vol. 4947, Springer.

[6] Ponzetto and Strube 2007. Deriving a large scale taxonomy from Wikipedia. *Proceedings of the national conference on artificial intelligence* vol. 22, no. 2, AAAI and MIT.

[7] Potthast et al. 2008. A wikipedia-based multilingual retrieval model. *Proceedings of ECIR '08*, Lecture Notes in Computer Science vol. 4956, Springer.

[8] Zesch et al. 2008. Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. *Proceedings of the Conference on Language Resources and Evaluation*, ELRA.

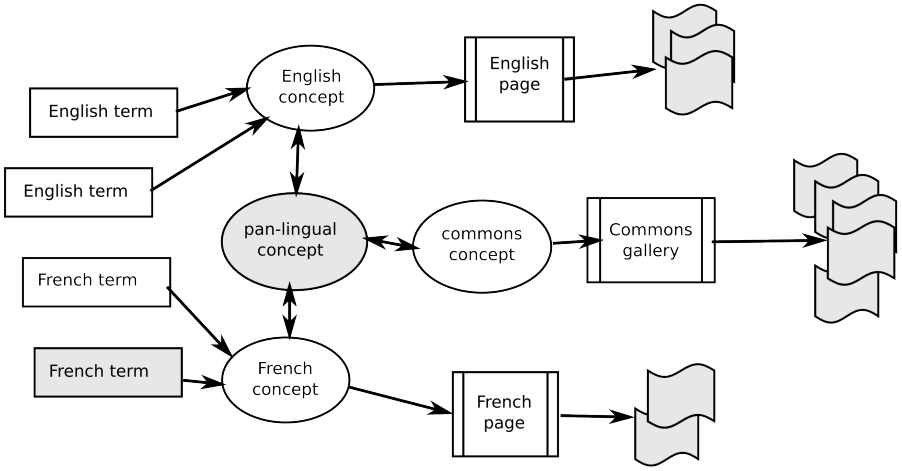

Fig. 1: A look-up path.