# Dynamically Generating Meaningful Links — Googlize.it!

**Willem Robert van Hage** and **Michel Klein**
Vrije Universiteit Amsterdam
AI / BI group, jointly known as the VU semantic web group
wrvhage@few.vu.nl, michel.klein@cs.vu.nl

## Abstract

Links are the kernel of the World Wide Web. However, up to now almost all links between webpages are hand-made. In this demonstration we show how knowledge on the web can be exploited to dynamically add links to webpages. We use explicit knowledge from ontologies and thesauri as well as implicit usage knowledge that is captured in the Google pageranks. The resulting application can be used as a proxy that adds new and unexpected but useful links to existing webpages.

## 1 Dynamically Generating Links

Currently, almost all links in webpages are hand-made. The author of a page decides at design-time for which terms he would like to add a link (i.e. the source of link) and he determines the target of the link. This is a useful scheme because it gives the author the control over the information that he would like to point his readers to. However, it also has its drawbacks: a person can only create links to information that he or she is aware of. As a consequence, the links in a particular webpage are restricted to the sources that are known by the author of the page. These are not necessarily the most useful or interesting links.

The World Wide Web itself contains a lot of information that can be used to suggest link candidates. In addition to the well-known web-directories that contain lists of relevant websites for specific subjects (e.g. the Open Web Directory[1], there is also other information that can be used to propose links, for example knowledge in thesauri, encyclopedia, ontologies and even knowledge about the popularity of pages on web. We have created an application that uses multiple online knowledge sources to dynamically add links to webpages.

## 2 Using External Knowledge Sources for Creating Links

When proposing links, there are two values that have to be determined: the source and the target of the link. For both we can use online knowledge sources. We distinguish two types of knowledge sources:

1. explicit encyclopedic information: knowledge structures that are created by humans, e.g. ontologies, thesauri, WordNet, Wikipedia;[2]
2. implicit usage knowledge: knowledge that can be derived from the global network of links on the web, e.g. Google pageranks.[3]

For determining the *sources* of links, i.e. the terms for which a link should be added, we can use knowledge sources of the first category. For example, we could add links for all terms in a webpage for which Wikipedia contains an entry.

For determining the *targets* of links, we can use information from both categories. If the ontology, thesauri or encyclopedia contains explicit references to webpages those can be used. In addition, we can use the category to which a term belongs to restrict the possible targets of the links. For example, for person names we restrict the link targets to homepages. If there is no explicit information about link targets, we can link to the page that has the highest relevance according to the Google-pagerank.

## 3 Implementation

We have created an application that functions as a proxy and adds links to webpages. The application is called "Googlize.it", or "Googlizer", referring to use of the Google-pagerank as mechanism to determine targets of links. While developing it further we add other sources of knowledge.

### 3.1 Overview

The initial implementation of the Googlizer uses the following algorithm.

**Parse** Parse the HTML of the page and extract all pieces of text. We accomplish this using the HTML::Parser Perl module.

**Tokenize** Using a controlled vocabulary we group words together that form a concept. We demonstrate this using the list of current Wikipedia entries.

**Recognize interesting concepts** Using a thesaurus we recognize a subset of these concepts that is we want to give special treatment. We demonstrate this with the Getty Thesaurus of Geographic Names (TGN)[4] for geographical names, which we link to a map showing their location on Map24. Something similar could be done for other classes of concepts.

[1] http://dmoz.org/

[2] http://en.wikipedia.org

[3] See for an explanation: http://www.google.com/corporate/tech.html.

[4] http://www.getty.edu/research/conducting_research/vocabularies/tgn/about.html

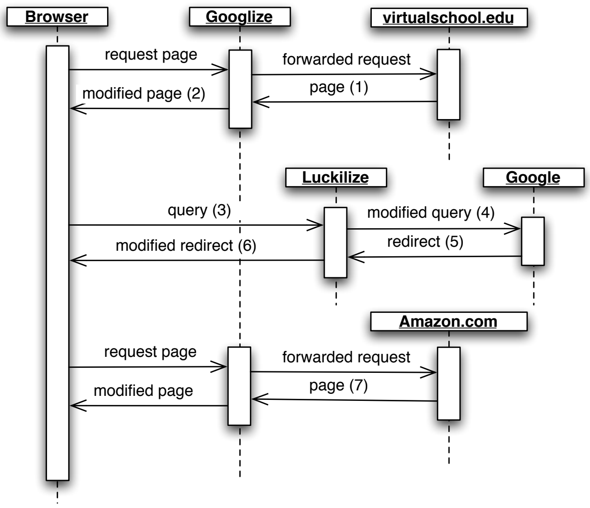


Figure 1: The default behaviour of Googlizer in a sequence diagram.

For example, names of people by using FOAF, or definitions of biomedical terms as demonstrated by the GOHSE system [Bechhofer *et al.*, 2005].

**Default behaviour** By default we link words to the Google "I'm feeling lucky" result page of the word and its direct context. The result page is also 'Googlized', which means that all existing links are changed to point to the Googlized version of the target page. This default behaviour is illustrated in Figure 1 and 2.

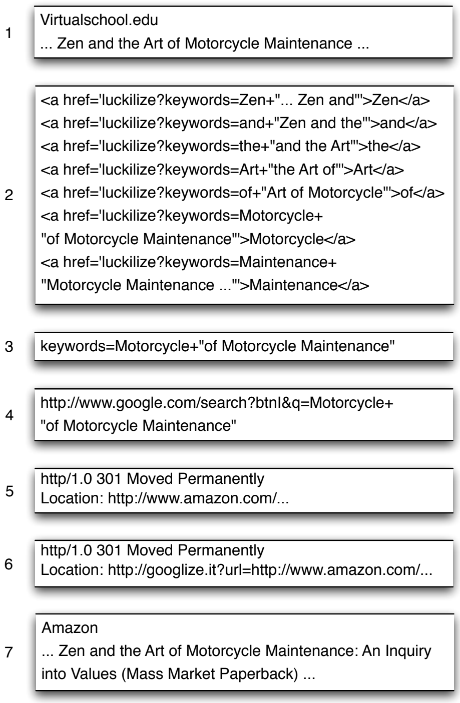
The figure describes a case where a user is reading a page about "Zen and the Art of Motorcycle Maintenance" on virtualschool.edu and he clicks on the word "Motorcycle" in the header of the page. The Googlizer sends him to the Amazon page of the book.

### 3.2 Googlize & Luckilize

The Googlizer system consists of two proxy servers: Googlize and Luckilize.

Googlize parses, tokenizes, and adds links to the HTML pages it forwards. It wraps every unlinked word, for example the word "Art" in the context of "… and the Art of … ", with a link to Luckilize. We provide Luckilize with the word and the direct context of the word. So `Art` is translated to `<a href='http://googlize.it/luckilize?keywords= Art+"the Art of"'>Art</a>`. This means that you can now click on any word in a page. Furthermore, Googlize prefixes every link target with the URL of the Googlizer. So `<a href="http://www.vu.nl">...</a>` is translated to `<a href="http://googlize.it/?url=http://www.vu.nl">...</a>`. This means every page you travel to will also be googlized. In order to keep the appearance of the translated page the same (as opposed to turning it into a 'Christmas tree' of underlined links) we add a "class=googlize" attribute to every added link that allows us to remove the link decoration using a CSS stylesheet.

Luckilize sends queries to Google's "I'm feeling lucky" service and changes the HTTP 301 redirect command it receives from Google to point to the Googlized version of that page. This causes the pages you travel to through the links Googlize added to be Googlized too.



```
1  Virtualschool.edu
   ... Zen and the Art of Motorcycle Maintenance ...

2  <a href='luckilize?keywords=Zen+"... Zen and"'>Zen</a>
   <a href='luckilize?keywords=and+"Zen and the"'>and</a>
   <a href='luckilize?keywords=the+"and the Art"'>the</a>
   <a href='luckilize?keywords=Art+"the Art of"'>Art</a>
   <a href='luckilize?keywords=of+"Art of Motorcycle"'>of</a>
   <a href='luckilize?keywords=Motorcycle+
   "of Motorcycle Maintenance"'>Motorcycle</a>
   <a href='luckilize?keywords=Maintenance+
   "Motorcycle Maintenance ..."'>Maintenance</a>

3  keywords=Motorcycle+"of Motorcycle Maintenance"

4  http://www.google.com/search?btnI&q=Motorcycle+
   "of Motorcycle Maintenance"

5  http/1.0 301 Moved Permanently
   Location: http://www.amazon.com/...

6  http/1.0 301 Moved Permanently
   Location: http://googlize.it?url=http://www.amazon.com/...

7  Amazon
   ... Zen and the Art of Motorcycle Maintenance: An Inquiry
   into Values (Mass Market Paperback) ...
```

Figure 2: The information exchanged between the browser and servers (numbers refer to the sequence diagram).

## 4 Future Work

Googlizer uses the context of terms and explicit and implicit knowledge to dynamically add links to webpages. However, because people differ the terms and their context are not enough for Google to decide what the correct target page is. Every user has a different perception of the world. Combining world knowledge with knowledge about the user could improve the subjective behavior of the Googlizer for each user.

For example, every instance of concepts that is of interest to a certain user could be linked to sources of information about the concepts. For example, if the user is a scientist names of articles could refer to the actual articles or CiteSeer[5] records about these articles, and names of other scientists could refer to their DBLP[6] record or their work homepage.

Even more futuristic, one could think of a googlized desktop. Everything on a computer could be infused with links. Clicking on a piece of text in an e-mail client should cause Googlizer-like behaviour and link to the most relevant information in your computer.

The current application can be found at `http://googlize.it`.

## References

[Bechhofer *et al.*, 2005] S.K. Bechhofer, R.D. Stevens, and P.W. Lord. Ontology driven dynamic linking of biology resources. In *Pacific Symposium on Biocomputing*, pages 79–90, 2005.

[5] Search engine and database of scientific citations, see `http://citeseer.ist.psu.edu/`.

[6] Manually maintained computer science bibliography, see `http://www.informatik.uni-trier.de/~ley/db/`.