

How to use the Web's information flood for teaching

Hermann Maurer
Institute for Information Systems and Computer Media
Graz University of Technology, Austria
hmaurer@iicm.edu

Heimo Mueller
Institute for Pathology
Medical University of Graz
heimo.mueller@mac.com

Abstract: In this paper we concentrate on the question how we can retrieve reliable information from the Web as source for knowledge in teaching and learning: The use of search engines has a number of dangers which together with potential remedies we will point out. Another approach to find useful information is to use "classical" resources of information like specialized dictionaries, lexica or encyclopaedias in electronic form. Some do suffer from what we will call the "wishy-washy" syndrome. It is interesting to note that Wikipedia which is also larger than all other encyclopaedias (at least the English version) is less afflicted by this syndrome, yet has some other drawbacks. We discuss how those could be avoided and present a prototype of a system that does take care of many of the problems mentioned, hence may be a model for further undertakings.

1. Introduction

The Web has turned into the dominant source of information. Most persons use it by employing one of the available search engines, or by going directly to a site they tend to rely on. Using one of the major search engines is tempting, yet one has to be aware of a number of problems: first, often more hits will be presented than anyone will ever look at; second, the reliability of information found is not at all guaranteed: it is up to the user to investigate whether results can be trusted or not; third, most current search engines are still based on a set of word in the query, rather than understanding natural language questions; fourth, the ranking of search results is not transparent and depends on many factors. We will examine those four points in Section 2.

Rather than using a search engine one might directly go to a specialized site. In Section 3 we discuss pros and cons of going to one of the "classical" sites. We exclude Wikipedia from this set on purpose since we will discuss issues concerning Wikipedia separately in Section 4. In Section 5 we present a prototype where an attempt is made to eliminate most of the problems discussed in previous sections. We argue that using the ideas mentioned we might finally get a very large repository of reliable material that we can base our teaching, our work and our judgements on.

2. Some aspects of search engines

One of the most obvious problems encountered when using search engines is that the number of search results is too large to be used systematically. Further, many search results are similar to others. Thus, we have an undesirably high degree of redundancy; worse, some search results are contradictory.

Concerning redundancy, it would be nice if search engines in the future would cluster together similar search results automatically, may be even combining results in a cluster into one more or less coherent document, so that users are only confronted with a limited number of clusters or documents representing the most important "views" on a topic. In some isolated cases this has worked quite well as was pointed out in (Wurzinger, 2010): it is shown there that in some cases redundancy can be cut by 75% using fairly simple similarity recognition algorithms as used for plagiarism detection, like in (Zaka B. et al, 2008), (Maurer H. & Kulathuramaiyer N., 2007); (Maurer H. & Zaka B., 2007) and (Kappe F. & Maurer H. & Zaka B., 2006).

To reduce redundancy dramatically (not by 75% but by 99%) and to retain coherent essays (i.e. to construct coherent essays from lots of tidbits that have been collected) is still science fiction today; yet it is one of the great research

challenges search engines are facing. At the moment, search engines could reduce the amount, and change the ranking of search results not just by “personalizing” them (as some already do), but to allow further searches in the set of results found. Recent attempts in this direction are search engines such as the slash-tags in (Blekko, 2010).

Concerning correctness, let us quote from (Wurzinger, 2010):

“We all accept that no information obtained is reliable (except if know we can trust the source of information), yet how dramatic the unreliability is can be shown with numerous examples. Searching for “boiling point of Radium” with Google two entries retrieved Aug.25 ,2010 are shown in Fig.1

The image shows a Google search interface. The search bar contains the text "boiling point of Radium". Below the search bar, it says "Ungefähr 58.400 Ergebnisse (0,13 Sekunden)". There are two search results listed:

- Chemical Elements.com - Radium (Ra)** - [Diese Seite übersetzen]
Name: **Radium** Symbol: Ra Atomic Number: 88. Atomic Mass: (226.0) amu. Melting Point: 700.0 °C (973.15 K, 1292.0 °F) **Boiling Point:** 1737.0 °C (2010.15 K, ...
www.chemicalelements.com/.../ra.html - Im Cache - Ähnliche Seiten
- Boiling Point > Radium** - [Diese Seite übersetzen]
The **boiling point of Radium** is 1140 ° C. Radium. Atomic Mass · Atomic Number. Boiling Point. Crystal Structure · Date Discovered · Melting Point ...
www.noblemind.com/search.exe?...Radium+Boiling+Point... - Im Cache - Ähnliche Seiten

Fig. 1: Boiling point of Radium.

One entry shows 1737 degree Centigrades, the other 1140. How should we know which one is correct?

May be life does not depend on this particular answer. However, consider a case we have been confronted with when we picked a type of wild mushrooms recently that we could definitely identify as “Gruenling.. When we checked its edibility we found five entries on the first search page, three telling us that it is a delicate edible mushroom, one informing us that it is deadly poisonous and one simply that it is poisonous!”

How is it possible that even in what seems reliable sources such wild discrepancies and contradictions can occur? There are two main reasons: one, often definitions differ: if you look for the “largest cave in Canada” do you mean largest by length, by volume, by height, or by which other criteria? If you want to know the height of a mountain on the moon do you mean the relative height compared to the deepest point “near” it, or do you mean the height above a hypothetical sphere giving the average height of the moon (as we sort of do on Earth when we compare heights to sea-level); second, discrepancies are often due to the fact that information comes from different times: it is very unfortunate that documents on the Web are rarely dated!

This, by the way, is the reason for the different judgement of the edibility of the mushroom mentioned above: it was eaten without known side effect for thousands of years; in 2002 suddenly two deaths seemed to be linked to the consumption of a dish made out of the mushroom. Those two isolated cases have caused new entries on the mushroom to call it poisonous when it might also be a kind of very rare allergic reaction!

What can be learnt from this: (a) if various definitions are possible the documents should make this clear: this is NOT a job for search engines but for authors of documents; (b) all documents should be clearly dated; (c) the date should be considered as part of the ranking algorithm in search engines. Note that if I search for ED-MEDIA I am likely to be more interested in more recent ones than in the ones ten years or further back!

Note also that search engines usually work with a group of input words, possibly connected by “or”, “and” or “not”.

A more linguistic approach was already taken in (Brockhaus 2006). Natural language queries have been allowed in this electronic dictionary now for over 5 years. One of the easy tricks was to observe the word at the beginning of the query: “Who” is clearly asking for a person, “Where” for a location, “Why” for an explanation“, etc. By delving more deeply into language understanding, Brockhaus ended up with fairly decent results.

Thus, most search engines are still based on words, albeit more and more cleverly. Inputting “Who was the inventor of the toothbrush?” into Google gives what it seems is a reasonable answer (“No exact date known...”). The search engine Bing finds William Addis in 1770 with a cute story: discrepancies within a search engine exist, discrepancies between different search engines can still be much more serious!

The question “Who was the physicist born in Vienna and died in Italy?” does not work well with Google. Since the search is text based, Google finds all Vienna physicists. Since Schrödinger worked in Italy at some stage his name pops up much earlier than Boltzmann. Bing actually finds Boltzmann better than does Google, and provides interesting further information, yet its search is also clearly word-based. In the prototype described at the end of this paper, documents have meta-data, hence Boltzmann is found easily.

However, general search engines have to work without systematic meta-data, so they either have to work with words or have to use deeper natural language understanding! Even if they do, how can we trust the result (see toothbrush example).

Summarizing: Major search engines do not yet employ deep language-analysis tools, are generally not good in allowing to narrow down large query sets and do not take enough care in reducing redundancy and taking dates in to account. Due to the importance of search engines further progress can be expected and emerging competition will be helpful.

3. Special purpose encyclopaedias and dictionaries

There are thousands of free encyclopaedias and dictionaries on the Web. Some give only limited access free of charge but ask for payment for “premium use” or such. Some (typically medical encyclopaedias) are only available for closed user groups (certified physicians). The (Austrian Encyclopaedia 1995) is one of the few that have survived till today. However, most famous encyclopaedias (e.g. in Germany Brockhaus and Meyer, the latter available online free of charge for many years, or the Britannica) have disappeared in printed form and are only offered with limited information for free, due to the pressure of free information, particularly from Wikipedia. A typical example is (Encyclopaedia Britannica, 2010) whose electronic premium version is quite remarkable.

Thus, Wikipedia has basically eroded the commercial basis of general purpose encyclopaedias. While this has been deplored by some critics like (Keen, 2007), claiming that this is the beginning of a road to mediocre material a vast number of persons believe that Wikipedia is such a valuable and also high quality tool that the demise of commercial products is acceptable. Although the authors of this paper have some points of criticism concerning Wikipedia they also are critical of traditional encyclopaedias for a reason often overlooked: the typical traditional encyclopaedia was an alphabetic arrangement of topics in an “objective“ way, thus reporting the “truth” about an event, a person, an idea, whatever. We believe such a concept is basically flawed. Everyone agrees that if we look at a material object (as sculpture, a mountain, a house, etc.) we can get a proper impression of the object only by seeing it from different views. This does also apply to non-material objects such as ideas, or personalities, etc., yet in general this is less explicitly acknowledged. But if we can only understand a complex person, a complex idea or a deep concept by getting many opposing views it does not help to present a single “compromise” or “wishy-washy” description of the issue. Rather, a number of different reports on the same subject with pointedly different views are needed. Traditional encyclopaedias have tried to live with this by having pro and contra views, yet there was always an author or team of authors behind each entry with a certain point of view, colouring the presentation. It is our belief that in future collections of encyclopaedic type this has to be avoided. Similarly, it has to be avoided that encyclopaedias present an issue from a single point in time, since this often hides important issues.

Let us explain this with one simple example. In the eighties of the last century Europeans were so worried about the extinction of interesting varieties of tropical wood that the import of certain types and objects made thereof was forbidden. A typical European encyclopaedia of 1985 would report this fact with some pride, showing the concern of Europe for maintaining variety in nature. However, since the import of tropical wood was not possible any more it lost its inherent value. Hence large forests of threatened species of tropical wood were burnt down to make room for rice fields that would yield at least a bit: The well-intended effort to protect tropical wood produced exactly the opposite of the desired effect. A European encyclopaedia of 2002 reported (a) that certain types of tropical wood are endangered and (b) that local population was continuing to destroy it. The reason for this was not mentioned.

This leads us to a critical analysis of Wikipedia. It turns out that Wikipedia might well be a step in the right direction, but that some changes would indeed increase its value still considerably.

4. Wikipedia

Wikipedia is certainly one of the big successes of the “Wisdom of the Crowd” paradigm as described in (Surowiecky, 2005, Wikipedia Foundation 2010).

Over time, many weaknesses have been pointed out: in addition to inadvertent errors there have been cases of deliberate spreading of false information including defamation of persons, blown out of proportion description by paid or unpaid fans of some notion or person, hidden advertisements, or discrepancies in numbers reported: In some report on a country the population of city A might be mentioned at a number x, while the report dedicated to city A might mention a different number y, potentially because census data from different time periods had been used. Another troublesome aspect is that the same event might occupy much space in some language version of Wikipedia, but may be quite short in other languages. Worse, the inventor of some device D might be person A in one country, and Person B in another country.

However, having said all this it is also clear that the average quality of contributions is quite good: the control of many readers is working to a high degree. Editing, censorship and correction procedures can be quite different in various language versions of Wikipedia. Rules are not carved in stone: in the English Wikipedia it was initially possible to write anonymously this has been given up. Today, at least some versions of Wikipedia do not allow writing contributions unless some screening of the writer has taken place before.

We have criticised that traditional encyclopaedias have only one entry for even the most complex topic, even if that topic can only be presented by presenting different points of view. Wikipedia is doing the same, yet it does allow to look at a thread of discussion that has lead to the current result, thus giving some additional insight.

However, we feel that a number of crucial improvements are still missing to make Wikipedia to what it is now trying to be: the ultimate source of reliable information on any subject whatsoever. To prove our point we are in the process of trying to reproduce what is good in Wikipedia, yet where the introduction of a number of additional features will help in achieving a new kind of quality. We have restricted the scale dramatically by only collecting information on a single small country and only issues of some stability involving it.

5. The Prototype

The system has been in operation since October 2009, covering items that involve Austria or Austrians in some way.

At the time of writing the prototype that can be tested at www.austria-forum.org comprises some 200.000 “objects”, an object defined as text-file, picture, audio- or video-file. The desired functionality and a first solid foundation information-wise is planned for completion by mid 2013, with some one million objects. It is important to understand the main differences between Wikipedia and the Austria-Forum:

(i) The domain is restricted to Austria as described, and it emphasizes information that has a high degree of stability. Thus, a biography of a former poet or the description of an event in history is well suited, a biography of a rising new star in politics is acceptable, results of the rescue of the Chilean miners or sports events of the last month have no place in Austria-Forum: “news type” information is left to the media. One reason is to avoid competition, the other is pragmatism: we cannot muster the resources to also cover all those items.

(ii) We distinguish between approved main entries and general entries in the community section. In the latter, rules similar to Wikipedia apply, yet contributions can be upgraded and moved to the main entries section if the editorial board so decides. Main entries have an author who has been screened and whose CV is available to users, so there is background information on who is writing what. Main entries are also taken from books and archives: in each case the aim is to provide a clear source.

(iii) It is our attempt to associate a date with each entry: not the last date of a minor update, but the date when the main entry was created. Note that this has two aspects: we hope to be able to e.g. show pictures with sliders that

allow to view the change of a city, a glacier, a river, etc., over time. Still more innovative and still in its infancy would be a slider that shows reports on some topic depending on the point of view.

The “time stamp” philosophy also means that if someone wants to do a major edit of an approved entry, this is not possible. Rather, a new entry with the same name is created. Thus, ideally, you will not find an entry on “nuclear energy” but a sequence of entries like “History of nuclear energy in Austria”, “Why nuclear energy is important”, “Why nuclear energy is dangerous”, etc: pointed and provocative contributions about nuclear energy from various points of view. Ideally, you should not find pictures of a city like Graz, or an essay about Graz, but photos of Graz at various times, and essays describing Graz at various times. Thus, quite in contrast to Wikipedia, an essay on Graz should no be updated, but retained as time capsule. Another time capsule will hopefully be added at some later stage.

(iv) Since contributions have a source and a date, it is possible to quote them in scientific contributions, an open issue with Wikipedia contributions. Austria-Forum is interactive in as much that anyone can add comments to a contribution: many comments may lead some editor to even write a new version of the essay, leaving the old essay with all its idiosyncrasies intact. Other communication facilities are also provided to hopefully strengthen the spirit of community.

(v) We do not believe in providing a single encyclopaedia, but a substantial set of them covering various topics. The reason is that the search in Austria-Forum allows to not only be narrowed down to one area (a very desirable feature) but to use metadata available. Note that Fig. 2 shows a form filled out with entries typical for a biography. Indeed the search finds immediately the person at issue (Boltzmann). But the form (metadata) required to find a lake, a building, a flower, would clearly have to look very different.

Suche in Biographien:

UND	Wien	Geburtsort	
UND		Geburtsland	-
UND		Geburtsjahr	- Jahr oder zwei Jahre mit - dazwischen eingeben
UND	Physik	Arbeitsort	-
UND		Arbeitsgebiet	-
UND	Italien	Todesort	-
UND		Todesland	-
UND		Todesjahr	- + Jahr oder zwei Jahre mit - dazwischen eingeben

Suchergebnisse für 'Geburtsort:Wien AND Arbeitsgebiete:Physik AND Todesland:Italien'

Seite	Relevanz
Boltzmann, Ludwig (Biographien)	100

Fig. 2: Searching in Austria-Forum using meta-data.

(vi) We have added to the Austria-Forum an new kind of object akin to an e-Book. Those books are stored in a kind of bookshelf: the first two rows with some historical books are shown in Fig. 3.

However, not only does the bookshelf look similar to a real bookshelf, also the books themselves behave more like real books than e.g. PDF-files, yet they do offer advantages like searches as one would expect from electronic substances. Books are also heavily cross linked with other information within the Austria-Forum and beyond. Allowing teachers and students to add their personal (or public) remarks and links will turn this new kind of object (Mueller, H. & Maurer, H., 2010) into a valuable teaching and learning tool. For readers eager to try this out look at e.g. <http://www.austria-lexikon.at/ebook/bookshelf/> and click at the first book on the shelf!



Fig. 3: Part of one of the bookshelves

6. Conclusion

Teaching and learning processes are going to increasingly use material on the web. In this paper we have analyzed ways how to retrieve reliable information. We have argued that no approach is without its flaws. We explained a substantial prototype that is currently developed that we hope will contribute to handling the flood of information.

References

- Encyclopaedia Britannica (2010). Electronic Version, <http://www.britannica.com/> last checked Nov. 30, 2010
- Wurzinger G. (2010). Data consolidation in large bodies of information; *JUCS* vol. 16, No. 21 (2010), 3314-3323
- Zaka B., Kulathuramaiyer N., Maurer H., Balke W.-T. (2008). Topic-Centered Aggregation of Presentations for Learning Object Repurposing; *Proc. of E-Learn 2008*, 3335-3342
- Maurer, H., Kulathuramaiyer, N. (2007). Fighting plagiarism and IPR violation: why is it so important?; *Information Services & Use*, vol.27, no. 4, IOS Press, 185-191
- Maurer H., Zaka B. (2007). Plagiarism - a problem and how to fight it; *Proc. of ED-MEDIA 2007*, 4451-4458
- Kappe, F., Maurer, H., Zaka, B. (2006). Plagiarism - A Survey; *J.UCS* 12, 8 (2006)
- Blekkko (2010). <http://blekko.com> last checked Nov. 23, 2010
- Brockhaus (2006). Der elektronische Brockhaus, Mannheim, Germany (2006)
- Encyclopedias (2010). <http://www.encyclopedia.com/> last checked November 25, 2010
- Keen A. (2007). *The cult of the amateur*. Doubleday 2007.
- Surowiecki J. (2005). *The wisdom of the crowds*. Anchor Books 2006.
- Wikipedia Foundation (2010). <http://wikimediafoundation.org/wiki/Home> last checked November 27, 2010
- Mueller, H., Maurer, H. (2010). A new approach for e-books for teaching and learning (to appear)
- The Austrian Encyclopaedia (1995). AEIOU. <http://www.aeiou.at>, last visited November 30, 2010.