Journal of Universal Computer Science, vol. 16, no. 21 (2010), 3314-3323
submitted: 3/9/10, accepted: 26/11/10, appeared: 1/12/10 © J.UCS

Information Consolidation in Large Bodies of Information

Gerhard Wurzinger
(IICM, Graz University of Technology, Graz, Austria
gwurz@iicm.edu)

Abstract: Due to information technologies the problem we are facing today is not a lack of information but too much information. This phenomenon becomes very clear when we consider two figures that are often quoted: Knowledge is doubling in many fields (biology, medicine, computer science,…) within some 6 years; yet information is doubling every 8 months! This implies that the same piece of information/knowledge is published a large number of times with small variations.
Just look at an arbitrary news item. If considered of some general interest reports of it will appear in all major newspapers, journals, electronic media, etc. This is also the problem with information portals that tie together a number of large databases.
It is our contention that we need methods to reduce the huge set of information concerning a particular topic to a number of pieces of information (let us call each such piece an "essay" in what follows) that present a good cross-section of potential points of view. We will explain why one essay is usually not enough, yet the problem of reducing a huge amount of contributions to a digestible number of essays is formidable, indeed is science fiction at the moment. We will argue in this paper that it is one of the important tasks of computer sciences to start tackling this problem, and we will show that in some special cases partial solutions are possible.

Keywords: information consolidation
Categories: H.3.4, H.3.5, H.3.7, H.4.3, H.5.1, M.1

1 Introduction

If we want to look for information on some particular topic one of the obvious ways to obtain it is to search for it in the Web. However, results are often of very dubious value as two well-known arguments show.

First, if we use one of the popular search engines we are bound to end up with such a vast number of hits that there is no way to systematically look at even a high percentage of them.

Second, if we find information how do we know it is reliable? 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. If we search for "boiling point of Radium" with Google the first ranked two entries as retrieved on August 25, 2010 are shown in Fig. 1.

Figure 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 picked a type of wild mushrooms recently that we could definitely identify as "Echter Ritterling" (Gruenling). When we wanted to check if it was edible or not 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!

The above two reasons have led to a number of attempted solutions, but most to not live up our expectation.

One such solution is not to look up information using a search engine but to rather consult some on-line encyclopaedia.

The one that immediately comes to mind is of course Wikipedia. There are, however, also serious problems with Wikipedia is. The most critical is that the author(s) of a particular contribution is (are) not known, nor is it known how up to date and (again) how reliable information found is.

In addition, and much less generally understood or discussed, we cannot expect Wikipedia to deliver strong points of view on issue where there are very diverging opinions. Wikipedia does sometimes include a discussion where more than one point of view is mentioned (like in the case of nuclear power [Nuclear power, 10]) but does not present altogether different views in separate contributions. Indeed in [Runciman, 09] it says explicitly "From its outset Wikipedia has aimed to operate according to a code of conduct (of which the centrepiece is the proposition that 'Wikipedia has a neutral point of view')". However, already in [Maurer, 04] it was pointed out explicitly that to get a complete picture, different points of view are needed. Indeed the argument is convincing and it surprising that it is not universally accepted. In a nutshell, the following analogy is invoked: if we take a physical object, e.g. a mountain, one cannot get a full understanding of that mountain by looking at one picture taken from one point of view; one has to look also at pictures taken from other points of view. What is true of physical objects is surely also true for non-physical constructs: one cannot understand a concept like god, or communism, or a political person, etc, if one gets only one point of view. Thus, one should not try to get wishy-washy contributions (results of compromises of various parties) in such cases but a number of contributions showing clearly different points of view.

If we do accept that often there is no absolute truth this seems to be the only viable alternative, an alternative that was even often neglected in traditional encyclopaedias. Sometimes even systematic quality control (missing in Wikipedia) is not good enough, one has to allow different points of view, as is e.g. done in [Austria-Forum, 10], a project we will concentrate on later.

There are numerous general purpose and special purpose encyclopaedias in addition to Wikipedia, with a variety of methods to improve the reliability of information. In Citicendium [Citicendium, 10] strict editorial quality control is enforced following the doctrine of Larry Sanger, once co-founder of Wikipedia resulting in high quality but small volume. There are hundreds of free encyclopaedias around, as you can verify easily using any search engine. [Encyclopedia, 10] claims to search 100 of them when a query is posted.

Former famous printed encyclopaedias like Britannica [Britannica, 10] or Brockhaus [Brockhaus, 10] are now available on-line [Britannica, 10], but only partly free of charge. There are innumerable special purpose encyclopaedias, some free, some for charge, some (particularly in the medical area) with access only to authorized persons. However, none of the solutions provided are free AND high volume AND reliable AND with consolidated information.

There are other ways of combining huge quantities of information. Some are based on gigantic digitization processes, with Google Books the largest and most aggressive one. There is also the approach to gather information by using portals. In Europe one of the most famous is the European Library [European Library, 10], a huge European project. Yet like all large portals it suffers from the difference in user interfaces. Also, although it provides massive amounts of information, it provides them in unconsolidated form.

2 Consolidating information in restricted information spaces

As was mentioned at the outset, what one really would like to get when investigating a particular topic on the Web is a small subset of relevant essays or collections of pictures with reliable quality. To achieve this on the Web in general is certainly not possible at the moment. The relevant question is: can it at least be done in some subsets of information spaces?

We believe that one can advance a bit in this direction, if the domain is restricted enough.

In [Afzal, 09], [Afzal, 10] and [Zaka, 08] and related papers a group of researchers have tried to collect scientific papers together that deal with a similar topic. Some of the techniques have been incorporated into the electronic journal JUCS [JUCS, 10] as is described elsewhere.

More interesting is that research that was done on plagiarism detection (similarity detection) like in [Maurer, 07a], [Maurer, 07b] an [Kappe, 06] can be applied to information consolidation to some extent.

Consider a fairly large (more than say 30) databases with textual material in a certain area of discourse. When asking a particular query a set of answers (usually at most one from each database) is obtained. Rather than returning all of them as answer, an attempt to consolidate the information can be made along the following heuristic lines.

We look at the ten or so longest contributions. Among them we determine first a set of contributions which by similarity detection techniques appear to be fairly different. We end up with a few contributions x(1), x(2), …, x(m), with m between 1 and 10.

Next we consider each other contribution y separately. We take junks (paragraphs) y(1), y(2),…., y(k) of y and again apply similarity detection techniques to find out for each y(t) whether it is similar to a part of one of the x(i).

The heuristics applied is now this:

Case 1: More than ½ of the y(t) is similar to some part of a specific x(i). All those parts are discarded, the other parts of y are added to that x(i). This is based on the assumption that y is basically a shorter version of x(i) with some extra ad-ons.

Case 2: If no x(i) can be found so that at least ½ of the y(t) are similar to a part of x(i) then all y(t) that are similar to some part of an x(j) are discarded, the rest is put in a set "other stuff". Here the philosophy is that y contains some material already contained in the set x(1), … x(m) (which we need not keep), but also some additional materials that are collected separately.

The answer to the query is the set x(1), x(2),…,x(m) where junks of information have been added, plus a further collection of junks, those representing the set "other stuff".

Note that the outcome are m reports that read coherently (the original x(i)) and have some odd pieces added to them at the end (junks of some y(t)'s), and further one collection of junks that do not read coherently at all, junks that have not been found in a similar form in any of the x(i).

It is clear that the parameters "at most ten x(i)", "½ of junks being similar to parts of a specific x(i)" , and the cut-off point when a junk is considered to be similar to some part of some x(i) are somewhat arbitrary. Yet in a reasonably large scale experiment the outcome was quite satisfactory. In a class on "societal aspects of computer science" at our university we had 20 topics, and collected 50 essays on each topic, encouraging students to study (and copy/paste from) contributions in the Web. With the technique described after consolidation not 50 essays for each topic would be obtained, but at most 11 contributions for each topic: at most 10 coherent essays followed by some add-ons, and one collection of "other stuff".

The results were encouraging. Essays were on average 2000 words long, i.e. for each of the 20 topics the 50 essays amounted to 100.000 words. On the average, the 50 essays per topic collapsed to only 6.3 "consolidated essays" with an average of 3.700 words, cutting the 100.000 words (50 times 2000) to about 23.310 words (6.3 times 3.700), i.e. to about ¼. Taking this together that there was noticeable loss in information, consolidation was quite successful. Mind you, since students used the best known search engines and probably copied extensively from some of the early resources they found it is not that surprising that a good percentage of pieces of text was more or less duplicating other pieces.

We will try to carry out larger experiment in the future by trying to consolidate information between 20 encyclopaedias. Some preliminary tests seem to indicate that the "consolidation factor" won't be as good, but we hope still significant.

3 Looking at one particular case

We will look now at a concrete example we have been involved in for some time. In this case the problem is not so much consolidation in the sense of compressing data, but more in the sense of collecting information concerning on a topic in one place or at least in a way that all pertinent information can be found easily. Our object of research is the so-called Austria-Forum [Austria-Forum, 10].

We start by discussing some important facets of the Austria-Forum and then explain the problems we wanted to solve.

The Austria-Forum is a rapidly growing collection of "Austriaca", i.e. documents that either have to do with Austria (with the region interpreted broadly, since Austria was much larger before end of WWI) or are/were done by Austrians. Material is in the form of text, pictures, videos, speech, music and other multimedia elements.

Information found in the Austria-Forum is supposed to be reliable. This is achieved in a number of ways.

First, it builds on the last printed Austrian encyclopaedia AEIOU (dated 2003) with the material now being updated.

Second, it has many contributions by specialists whose name as author is explicitly shown with a CV on the server, each contribution with the date it was written or when a major update occurred.

Third, it is based on well established archival material from various large libraries, museums and archives in Austria, again each source quoted.

Fourth, it is based on books, again with authors, titles, publishers and dates of publications shown.

Thus, some information may be outdated: but it is shown when it was composed, and at that point it was quality information. Also, although prominent biographies of living persons, or up-to-date information on cities are included, no attempt is made to present any news items, certainly not if they will not be of interest a few years. However, contrary to an encyclopaedia published at a certain time it is not intended to present a time slice, but rather also the development over time. Hence, ideally, e.g. population figures for a city are not shown for today, but the development of population over time. In addition to a picture of some mountain valley now, the Austria-Forum attempts to also show what that mountain valley looked like 50 or 100 years ago, etc.

Another distinguishing feature is that the Austria-Forum is not one encyclopaedia, but a large (some 50) specialized encyclopaedias in addition to a general one. The rationale behind this is that for searching, often meta-data is necessary, and the meta-data will be different for biographies, flowers, animals, minerals, lakes, etc.

An example will best explain the situation. If we are looking for the person "Kreisky" a simple search will do. But what if we want to find the famous Austrian researcher in physics, whose name we can't remember, but we can remember that he was born in Vienna and died in Italy. In such cases usual search engines may fail. Fig. 2 shows the kind of form we fill out in such a case. It also shows the search result: Ludwig Boltzmann. The initial part of his biography is shown in Fig. 3.

Figure 2: Searching via meta-data

Figure 3: Biography Ludwig Boltzmann

Before we now discuss one of the problems of information consolidation that was successfully tackled it is worth-while mentioning that in Austria-Forum we hope that contentious issues will result in contributions form various points of view. Although this and the other aims mentioned are partially met by the current version of Austria-Forum (with the respectable number of some 150.000 entries). Austria-Forum is work in progress and the project manager Hermann Maurer [Maurer, 10] is planning to finish the groundwork only by Qu. 1, 2013. Why one can talk about finishing the groundwork at all is because the majority of entries will not have to be checked any more, the need to re-check or not is a particular meta-data entry (the description of an Austrian poet who has already passed away will not change, nor will historic facts, nor will the flora change dramatically).

One typical problem of the Austria-Forum, made particularly difficult due to the fact that we are concerned not with one encyclopaedia but with a set of them is that e.g. a picture of a city maybe found in the contribution of the city, or in the encyclopaedia of pictures of Austria, or in the encyclopaedia of castles or castle ruins

(of which we have over 2000 in Austria), or as part of a book about some region or period of history of Austria, etc.

Hence one type of consolidation that we had to tackle was to assure that if we look up an entry of a village or city in Austria, all relevant information is made available at our fingertip, no matter where that information is located. Since different villages may have the same name (there are 9 Neuhofen in Austria, 11 St. Martin, etc.) proper collection of data is less than trivial. Using a set of heuristics that was sure to deliver the right solution in 99% of the cases, but was set up so that in the 1% remaining cases a human could solve the dilemma a first success in consolidation was to connect contributions from the general encyclopaedia to the picture encyclopaedia, and conversely. Fig. 4 shows part of the entry for the city of Melk. At the end of the entry there is a link to the corresponding picture sequence in the Encyclopaedia of pictures of Austria as shown in Fig. 5 automatically generated by our heuristics. Indeed thumbnails are shown. Each picture can of course be enlarged.

Figure 4: Encyclopaedia: Melk

Figure 5: Encyclopaedia of pictures: Melk

At each moment a return to the main encyclopaedic entry has been provided by the heuristic approach hinted at, a link to "AEIOU", the trademark for the main encyclopaedia.

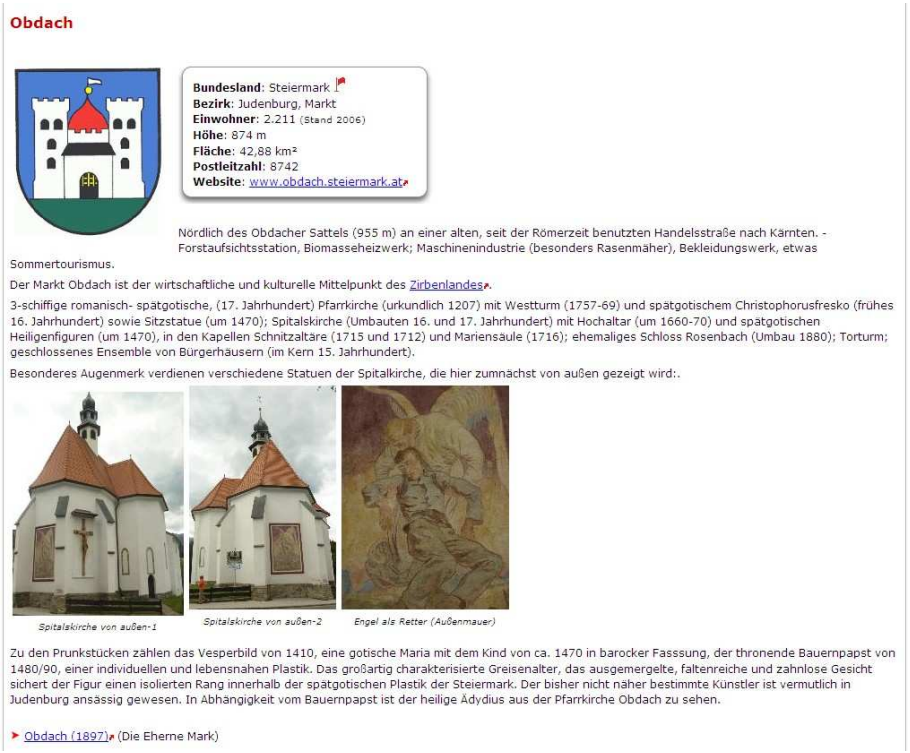

Figure 6: Encyclopaedia: Obdach, 2010

The consolidation has to cross time boundaries. A picture of one of the famous churches in the town of Obdach taken in 2010 as shown in Fig. 6 leads to a view of Obdach in 1897 in Fig. 7 with the church.

394 Obdach.

In zahlreichen Vorträgen, in welchen Falb nach seiner Art Religion zu lehren, seinen Schülern vielfach die Größe der Schöpfung durch Betrachtung des Sternenhimmels zeigte, machte sich immer bestimmter der mächtige Zug Falbs zur Astronomie geltend. Die Stellung als Erzieher in einem adeligen Hause, die Falb bald hierauf antrat, verschaffte ihm eine unabhängige Existenz. Falb gründete in dieser Zeit die populär-astronomische Zeitschrift „Sirius", welche er bis 1877 dirigierte. Bald darauf publicierte er das Werk: „Grundzüge einer Theorie der Erdbeben und Vulcanausbrüche", worin er zum erstenmale seine Theorie über den Einfluss der Mondnähe auf die Erdbeben-Erscheinungen aufstellt. In den

Obdach.

Figure 7: Encyclopaedia: Obdach, 1897

Fig. 7 shows the old book in the "book-mode" available in the Austria-Forum. The little elliptical icon indicates that there is also a modern panoramic view of the town available.

4 Summary

In this short paper we have introduced what we consider one of the major issues in distributed information systems in this century: information consolidation. We feel that the necessity of new techniques to reduce information to what is essential on the one hand, and to pull relevant information together on the other hand has not been stated as explicitly as we have tried to do in this paper. We have only shown two simple examples. Currently, another Ph.D. student in our group is involved in a major undertaking involving information consolidation: the Austria-Forum has obtained access to the largest picture data-base of Austria with over 10 million pictures, growing daily by another 4.000. It is a tremendous challenge to try to incorporate about 1% of those pictures (100.000) so that existing ones are not duplicated and that only those pictures relevant to the aims of the Austria-Forum are selected. It will be another large set of heuristics that will do a high percentage automatically, yet will leave the decision whether and where to incorporate a picture to a human in case of doubt.

Acknowledgement

I am indebted to Professor Maurer for introducing me to the topic of information consolidation and allowing me to use some of his personal data (from the student project) for this paper.

References

[Afzal, 09] Improving Citation Mining: M. T. Afzal, W. T. Balke, N. Kulathuramaiyer, H. Maurer; Proc. of International Conference on Networked Digital Technologies (NDT 2009) Ostrava, The Czech Republic, July 29 - 31, 2009. pp. 116-121. Publisher: IEEE.

[Afzal, 10] Rule based Autonomous Citation Mining With TIERL: M. T. Afzal, W. T. Balke, N. Kulathuramaiyer, H. Maurer; JDIM (Journal of Digital Information Management), Volume 8, Issue 1, 2010, pp. 196-204

[Austria-Forum, 10] Das Austria-Forum, http://www.austria-forum.org, checked August 25, 2010

[Britannica, 10] Britannica: http://www.britannica.com/, checked August 25, 2010

[Brockhaus, 10] Brockhaus: http://www.brockhaus-enzyklopaedie.de/, checked August 25, 2010

[Citicendium, 10] Citicendium: http://en.wikipedia.org/wiki/Citizendium, checked August 25, 2010

[Encyclopedia, 10] Encyclopedia: http://www.encyclopedia.com/, checked August 25, 2010

[European Library, 10] European Libary http://search.theeuropeanlibrary.org/, checked August 25, 2010

[JUCS, 10] JUCS: http://www.jucs.org, checked August 25, 2010

[Kappe, 06] Plagiarism - A Survey; F. Kappe, H. Maurer, B. Zaka; J.UCS 12,8 (2006)

[Maurer, 04] Der Berg von hinten; in: H. Maurer, Der Anfang, XPERTEN, freya pub.co. Austria (2004), pp.176-177, see also http://www.iicm.edu/XPERTEN

[Maurer, 07a] Fighting plagiarism and IPR violation: why is it so important? H. Maurer, N. Kulathuramaiyer; Information Services & Use, vol.27, no. 4 (2007), IOS Press, 185-191

[Maurer, 07b] Plagiarism - a problem and how to fight it; H. Maurer, B. Zaka; Proc. of ED-MEDIA 2007, 4451-4458

[Maurer, 10] H. Maurer: http://www.iicm.edu/maurer, checked August 25, 2010

[Nuclear power, 10] The nuclear power debate, http://en.wikipedia.org/wiki/Nuclear_power_debate, checked August 25, 2010

[Runciman, 09] Review of the book "Andrew Lih, The Wikipedia Revolution: How a Bunch of Nobodies Created the World's Greatest Encyclopedia, Hyperion (2009): Like Boiling a Frog, David Runciman, London Review of Books, Vol. 31, No. 10 (May 2009), pp. 14-16

[Zaka, 08] Topic-Centered Aggregation of Presentations for Learning Object Repurposing; B. Zaka, N. Kulathuramayer, H. Maurer, W.-T. Balke, Proc. of E-Learn 2008, 3335-3342