Producing Unifying Reports on Topics of General Interest

Delilović, Namik and Maurer, Hermann

Abstract: We try to argue in this paper that the WWW is full of large collections of information, yet the reliability is always in doubt, independent if the information comes from a company server, an institution that was usually founded for some specific reason, or even a system like Wikipedia: certain aspects are often omitted, distorted or not dealt with enough. In this paper, first attempts and ideas on how to organize such a large body of very diverse material in a way to assure high quality and reliability are discussed. We also present some preliminary but encouraging results.

Index Terms: Digital Information, Reliable Information, Plagiarism, Automatic Linking, Knowledge Discovery

1. INTRODUCTION

Paraphrasing part of the Preface of the Report by Naja BENTZEN (Policy Analyst, Directorate-General for EU Parliamentary Research Service) we have:

"*The danger is that we are moving from a rational (fact-based) society to emotional (not really fact-based) society since one cannot believe the wave of contradictory information on the Internet*".

2. MAIN IDEAS AND ISSUES WITH OUR APPROACH

For the above reason, it is important to carry out research and experiments on how to group reports that somehow belong together, to get an overview of what is happening in certain areas and to minimize the danger of fake news and half-truth spread via social networks. A first implementation cannot be done on the full WWW as it now exists. However, methodologies have to be found by research or have to be invented and experimented with on substantial multi-faceted material with results that hopefully will prove helpful also in lager contexts than the few million documents we propose to use as experimental ground.

The main idea is to collect for many topics of interest lists of contributions dealing with the subject but also allowing to prioritize them by date, length, quality of origin, the sentiment expressed and other parameters that can be chosen.

This requires to define a set of metadata that can be extracted from the body of material (be it WWW pages, digitized books, other information files in several formats (like PDF and WORD) or multi-media material enriched with metadata.

Besides, discussion forums should be available and open to anyone (it makes sense that users can use pen names. Their email addresses are encrypted, and only known to the system (in case legal matters arise). However, each forum has experts (shown as such) reporting only reliable facts, and able to correct, add to or delete dubious statements, whether they were put there intentionally or not.

We will discuss work we have started in [6] in the following sections.

3. CONCENTRATING ON A LARGE YET TYPICAL SUBSET OF THE WWW

As explained, we are first applying our ideas and algorithms to a specific yet large collection, consisting of all kinds of material. To be specific, we concentrate on the 1.2 million media objects in austria-forum.org [21] for a number of reasons:

(i) Documents in this collection come in many formats: as wiki/html pages, as word- or PDF files, in some picture or video format, or some web-books style (two different ones are in use in just Austria-Forum (AF for short from now onward) necessitating approaches that are fairly independent of data format.

(ii) We have made first attempts to find contributions that somehow belong together. As an example, if we choose any of the topics under [4] we will find (a) a list of companies, mostly stationed in Styria (the

province that is mainly responsible for AF) that have a connection with the topic chosen and (b) a list of hopefully most contributions in AF that deal with the topic. Part (a) is simple since the information can be extracted from existing databases; part (b) is tricky: we developed a semi-automatic approach that for each topic in "relevant parts" of AF searches for likely candidates are carried out, the result presented as a list considered to contain all suitable candidates. This is done with automatic methods that will need refinement. However, all not fitting are ignored manually, the others are automatically added to the list.

(iii) This sounds easier than it is: Note first that we have put "relevant parts" under quotation marks, since how can they be determined? Well, when starting AF, we decided not to use an encyclopedic (alphabetic) approach but to introduce different (semi - hierarchical) categories and assign reports etc. accordingly; even this is not always easy; hence we use overlapping-hierarchies to be able to assign a report to more than one category. However, this means that, e.g., contributions on agriculture collections checking in historical music collections can probably be omitted. Secondly, even if the system makes many suggestions for reports that look suitable, a person has to make the final decision. Nevertheless, the person cannot spend the time to evaluate each report in detail. Hence methods need to be developed to speed up this decision making a lot. There are also completely different approaches like [30].

(iv) Another reason to use AF as experimental ground is that AF contains hundreds of thousands of non-textual materials: we have started to examine pictures. However, music or video-clips etc. will eventually also have to be handled. Concerning pictures, in order to group them together, the apparent idea is to use titles, metadata (such as the time when pictures were created) and picture comparison techniques. None of this works well [29]. In essence, metadata come in different formats, titles vary a lot, and picture recognition does well with very similar pictures but to recognize that, e.g., two photos show the same person when they were taken 30 years apart is more than a challenge. On the other hand, the relatively recent advances in automatically making a young person many years older show that even such tasks can be (partially) tackled. Overall, the merging of somewhat different image databases can be tackled to some extent, see, e.g. [1].

(v)Work on AF started in 1995 (!), on the Internet seriously a few years later, and under the name AF officially 2009. It soon became apparent that information on the Web was not reliable in many ways. When it came to checking numeric data, the idea of checking values, not in one but many databases proved quite helpful. It came as a big surprise that even "obvious" answers like "How large is continental France" cannot be answered [2]. Indeed using [10] you find six different values in six different sources:

Factbook:	643801
DBpedia:	674843
Geoname:	547030
Infoplease:	547030
Britannica:	543965
Wolfram:	551500

Once we realize that the first two large numbers result from incorporating overseas departments French Guiana, Guadeloupe, Martinique, Mayotte, and Reunion the differences are only moderate but still leave us uneasy.

(vi) Indeed, just a bit of more checking leads to an awful truth: When we use a search engine to find information on the WWW, we often find answers, but without definition on what the answers are based on. However, what does it help us to find Mulu as the largest cave system in SE Asia, when we are not told, largest in what sense: By volume? By the length of corridors? By the largest height of a part of the cave? Thus, the situation is already very complicated when looking at geographic parameters. Consider a trivial example: Do we measure the size of an island at high or low tide? Surprisingly, geographers do not agree on a definition; nor do they tell us when a river is still part of the country it flows through or when it is already part of the sea. Since salinity changes because of the tides, a definition would be good to have, but there is no universally accepted one! [3].

(vii) To make information, providers recognize how often definitions are missing, we allow simple ways of feedback [5] and want to mention a few more typical examples. Is it not surprising that Lhotse is considered in most sources as the 4th highest mountain on earth? However, if we look at the data, it is in the same ridge as Mt.

Everest, "just" separated by a crack of 600 m depth. This information is enough for most geographers to call Lhotse a separate mountain and not a side-peak of Mt. Everest. However, if we check for high mountains in France, we find surprisingly many. The reason is that in France a crack of 200 m depth is sufficient to define a new mountain, rather than just to consider it a side-peak of the higher mountain!

We do hope that we can contribute with this example that definitions are accepted internationally since the Internet as an international affair would require this (for more such examples see [32]). Furthermore, if there is no standard, the net should give us not only an answer but also the definition on which the answer is based on.

Realizing that even numerical facts in the WWW are unreliable since the definitions, on which the answers are based on, are missing. Therefore, we have come to two conclusions: One, we want to try with this paper to convince many information-providers to specify the definitions they work with; and two, we want to expand our attention to reports dealing with critical issues of our society by trying to present an overview of opinion on such matters and why some remarks can be taken more seriously than others.

(viii) It is clear that checking numeric data is still "reasonably" easy, but to find pieces of textual data that deal with similar topics will require a mix of modern techniques from language processing, AI, style/plagiarism detection and others. We mentioned only half of the problem: when finding material on similar subjects which ones can we trust? A sophisticated combination of origin of the document, time when it was written, automatic sentiment analysis and others will allow first approaches to this challenging problem. In the end, it will again be trusted experts who will support one group of statements or another. We can only hope that experts do not come too often to different conclusions. However, by identifying experts who systematically take a different point of view even in "clear" cases, the selection of experts can be improved continuously in a machine-learning fashion.

4. A First Step: Rough Grouping of Documents.

In a large universal system like WWW or a subset like Austria-Forum documents will range from new technologies to reports on sports events, from art to historical information (of any time-period and most areas of the world), from nature to geography, from medical information to political one, etc.

Ideally, documents should have metadata associated with them as has been suggested over and over again; however, no universal model has ever been accepted. Typical types of metadata may include:

(i) Description of type of content like in a classification system. The well-known DDC (Dewey Decimal Classification system) first published in the United States by Dewey in 1876 defines main categories often used in (ordinary) libraries:

000 – Computer science, information & general works
100 – Philosophy & psychology
200 – Religion
300 – Social sciences
400 – Language
500 – Pure Science
600 – Technology
700 – Arts & recreation
800 – Literature
900 – History & geography

Each category comes with two further levels of subdivisions. Many variations of the system have been introduced. Other typical entries from the DDS are: *means of creation of the data, purpose of the data, source of the data, process used to create the data*.

Note that in AF the structure and metadata vary from category to category:

In biographies the typical entry is:

[{Metadata Geburtsort='Zamberk' Geburtsland='Tschechische Republik' Geburtsjahr='1841' Arbeitsgebiete='Chirurg, Medizin' Todesjahr='1900' Todesort='Zamberk' Todesland='Tschechische Republik' Suchbegriff='Chirurg Medizin ' Kontrolle='Nein'}]

In essays a typical entry may be much simpler and not as well structured:

[{Metadata Suchbegriff='Eric Kandel, Neurowissenschaften, Demenz, Spazierengehen, Erinnerung' }]

Of course, the W3C consortium is also deeply involved in the matter of metadata and related to it, the semantic web [7] and the Sematic stack Platform [8].

Conflicting interests of researchers and many industrial partners involved have made it impossible to come up with a generally excepted solution with the possible exception of the HTML5 protocol [9].

It is a pity there is nobody really in control of WWW and desirable future features, so we have taken a conservative approach in our attempts:

We provide each document with a description according to the DDC classifications. This has received much support from the library community, an essential part of our activities, see [41]. Moreover, if possible, we are also taking care of some items mentioned under BBC, particularly:

The creation date of the data, the purpose of the data, source of the data, the process used to create the data.

As other inputs, we are using the title of the document (and its subtitles in the documents, if provided.) Note that much care has to be exercised. Many of documents are written by top journalists, who choose titles that have nothing to do with the content of the document to attract readers or make them curious.

5. Grouping Reports into Topics: The Crux of the Problem: what Technologies are Available?

In (4), we grouped documents in very different groups. In this phase, the question is now how to refine the documents into groups, so they belong to some specific topic.

Accepting that metadata classification schemes are not enough to collect all relevant information on a topic, we use four different techniques to find "documents that somehow deal with a similar topic."

(i) The obvious approach is to use a word-based approach based on a dictionary with stop-words removed and synonyms introduced, and such. This approach has been successful in simple cases as e.g., in [12], [24], [26] or [25]. However, instead of using just words, larger pieces of text will increase accuracy, yet require more clever techniques. Some such techniques have originally been used for plagiarism detection to identify if a piece of text is similar (copied from?) some other source.

(ii) A rich repertoire of plagiarism detection methods has been developed with early treatments already in [18], [19], [22] and [23] and have much been extended in recent years, like, e.g, in [11], [14], [15] or [17]. An up-to-date systematic examination of plagiarism detection techniques is [16]. Much of this is based on better and better language-understanding techniques. [27] is a good survey and source to many up-to-date researches in this area. An often-quoted older paper is [28] since it also applies to non-textual data: Basically, it uses "Term Frequency Inverse Document Frequency", obtained by comparing the frequency of words or phrases in a document with their frequency in average material. Other ways to detect similarities using linear segmentation models such as in [12] and [13] have also been used successfully.

(iii) Current up to date mechanisms are using machine learning, particularly what is often called *Deep Learning* and *Word Embedding Methods*.

(iv) Note that other aspects like style-similarity in certain segments are both valuable for plagiarism detection but may also show that some documents need not be considered since they are derivatives of others.

6. How to Involve Everyone Who Wants to Be Involved (Including Troublemakers)

It is the intention to also use the system with its reliable contributions very much as a powerful weapon against half-truth and fake news.

For this purpose, as mentioned before, the system will allow anonymous discussion forums that are supervised by one or more experts.

To be specific, suppose the European parliament (or a national parliament) passes a new law, recommendations or changes thereof.

Right now, there are even paid persons who on purpose subtly distort the decision and spread it all over social media to influence people's feeling. In the future those anonymous contributions are welcome, but at some stage will be followed by a marked passage containing an accurate quote of the parliamentary decision (possibly provided in somewhat easier to understand form if the matter is complicated and requires references to other decisions). Also, an explanation of why the decision was taken (possibly even showing

figures for the pro and contra votes) might be supplied. Clearly, this can only be done by a reliable person (like a journalist) attending some of the meetings and would not apply to confidential matters.

There are also technical developments that should be discussed openly but with the trustworthy contributions of experts.

A typical example might be electric cars. It is clear that fully electric cars based on Ion-Lithium batteries decrease pollution where they are used but increase pollution (CO2) where Lithium is produced. In terms of CO2 output, worldwide usage of fully electric cars is most likely not beneficial. The situation starts to look different if new types of batteries can be developed or if certain types of hybrid cars are used, e.g, cars that drive short distances (up to, e.g. 50 km) only on electricity, while the combustion motor is only used to generate electricity for longer distances. Then the question is: From where do we get clean electricity? It might be worthwhile to have an open discussing how nuclear energy is doing compared to electricity generated by photovoltaic elements, solar boilers, winds, water, tidal energy generators, methanol technology [20]. What alternative do we have to Uranium (Thorium Saltwater reactors that are as powerful but much less dangerous than Uranium reactors. How about fusion? Will such reactors ever be economically feasible?).

A note may be appropriate here: How come if Thorium Reactors are such good alternative to Uranium reactors, not much research has been invested in them? The answer is sobering: all states original interested in Uranium reactors, USA, Russia, France, UK, India, Pakistan, China and Israel did not intend them to produce energy but to produce atom bombs. And for this end Uranium is better than Thorium; hence Thorium was largely ignored.

7. Other Classifications

The idea is to extract metadata hidden in a document, also using style and sentiment analysis to allow users when searching to specify one of many preferences; this should, for example, include: date of publication, source of publication (reliability), length of publication, standing of authors according to h–index on Google Scholar [31-35] or such, altogether allowing users to find out that despite conflicting points of view some are dominated by persons having significantly higher credibility, thus trying to provide a tool to distinguish real from fake news.

Note that the source is particularly important. Have the authors published in a cheap daily newspaper or in a book by a known high-quality publishing company? Are they particular qualified by distinctions as scientists or high public function? Do they have particular commercial interests to support a certain point of view, and similar?

8. Aim of the Effort

To continue our attempts at solving the tremendous number of issues when it comes to the problems of reliable knowledge gathering of which some have been pointed out by encouraging the community to work on such topics and possibly with us. Let us be clear about one issue: "We should neither discuss the advantages nor disadvantage of digitization, digital libraries, WWW and such. Arguments abound in both directions. We should start to control and define what our digitized future should look like, and not leave this to chance, commercial interests or billionaires".

9. Conclusion

We have tried to show in this paper first attempts and additional ideas on how to organize a large body of very diverse material in a way to assure high quality and reliability. This is not a small undertaking, so we will not be able to reach our aims unless we have been able to convince part of the community to work with us in the mentioned directions.

References

[1] R. Mehmood, H. Maurer: Merging image databases as an example for information integration; Central European Journal of Operation Research, vol.23, no.2 (2015), 441-458

[2] R. Mehmood: Geographic data verification; IPSI BgD Transaction son Internet Research, vo.10, no.2 (2014), 20-25

[3] M. Kulaturamayier, H. Maurer, R. Mehmood: Some aspects for the Reliability of information on the Web; JUCS, vol.20., no.9 (2014), 1284-1303

[4] https://austria-forum.org/af/Thema

[5] Namik D, Hermann M (2019) A Note Concerning Feedback and Queries for Web Pages. Journal of Universal Computer Science 25 (7):733-739

[6] Maurer H, Delilovic N, Zaka B (2019) Libraries of Interactive Books as Powerful Tool for Information Communication. Paper presented at the EdMedia + Innovate Learning 2019, Amsterdam, Netherlands,
[7] https://www.w3.org
[8] https://www.w3.org/standards/semanticweb/
[9] https://en.wikipedia.org/wiki/HTML5
[10] http://austria-forum.org/af/Geography/Europe/France/Geography
[11] Muhr, M., Kern, R., Zechner, M., & Granitzer, M. (2010). External and Intrinsic Plagiarism Detection using a Cross-Lingual Retrieval and Segmentation System Lab Report for PAN at CLEF 2010. In 2nd International Competition on Plagiarism Detection.
[12] Kern, R., & Granitzer, M. (2010). German Encyclopedia Alignment Based on Information Retrieval Techniques. In M. Lalmas, J. Jose, A. Rauber, F. Sebastiani, & I. Frommholz (Eds.), Research and Advanced Technology for Digital Libraries (pp. 315–326). Springer Berlin / Heidelberg. http://doi.org/10.1007/978-3-642-15464-5_32
[13] Kern, R., & Granitzer, M. (2009). Efficient linear text segmentation based on information retrieval techniques. In MEDES '09: Proceedings of the International Conference on Management of Emergent Digital EcoSystems (pp. 167–171). http://doi.acm.org/10.1145/1643823.1643854
[14] Nanda, R., Siragusa, G., Di Caro, L., Boella, G., Grossio, L., Gerbaudo, M., & Costamagna, F. (2019). Unsupervised and supervised text similarity systems for automated identification of national implementing measures of European directives. Artificial Intelligence and Law, 27(2), 199-225.
[15] Rexha, A., Kröll, M., Ziak, H., & Kern, R. (2018). Authorship identification of documents with high content similarity. Scientometrics, 115(1), 223–237. http://doi.org/10.1007/s11192-018-2661-6
[16] Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: a systematic literature review. ACM Computing Surveys (CSUR), 52(6), 112.
[17] Bassa, A., Kroll, M., & Kern, R. (2018). GerIE - An Open Information Extraction System for the German Language. Journal of Universal Computer Science, 24(1), 2–24.
[18] Maurer, H.; Zaka, B., Kappe, F.: Plagiarism – A Survey: JUCS, vol. 12, no 8 (2006), 1050-1084.
[19] Kolbitsch, J., Maurer, H.: Community Building around Encyclopaedic Knowledge, CIT vol.14, no. 3 (2006), 175-190.
[20] Weizäcker, E.U., Radermacher, F.J.: http://austria-forum.org/af/Wissenssammlungen/Essays/Wirtschaft/Methanol_Wirtschaft
[21] http://austria-forum.org
[22] Afzal, M-T.; Kulathuramaiyer, N.; Maurer, H.: Creating Links into the Future, JUCS, vol 13, 9 (2007), 1234-1245.
[23] Kulathuramaiyer, N.; Maurer, H.: Fighting plagiarism and IPR violation: why is it so important? Learned Publishing, 20, no. 4, (Oct. 2007) 252-258
[24] Afzal, M.T.; Kulathuramaiyer, N.; Maurer, H.: Expertise Finding for an Electronic Journal, Proceedings of I-KNOW and I-MEDIA 2008, 436-440.
[25] Afzal, M.T.; Balke, N..; Kulathuramaiyer, N.; Maurer, H.:Rule based Autonomous Citation Mining with TIERL; Journal of Digital Information Management, vol. 8, no 3 (June 2010), 196-204.
[26] Maurer, H.; linear ordering of a multi-parameter universe is usually nonsense, H. Maurer, Theoretical computer Science, Volume 429 (2012), pp. 222–226
[27] http://en.wikipedia.org/wiki/Natural_language_processing
[28] Smihtz, M.A.; Kanade, T.: Characterization through the combination of Image and Language Understanding, Carnegie Mellon University 1997 (CMU-SC.97-111) [29] Glatz, M.; Maurer, H., Afzahl, M.T. Finding Reliable Information on the Web Should and Can Still Be Improved CIT- Journal of Computing and Information Technology vol.26 no. 1 (2018), 1-6.
[29] Haider, S.; Afzal, M.T.; M. Asif, M.; Ahmad, A.; Abuarqoub, A.; Mauer, H.: Impact analysis of adverbs for sentiment classification on Twitter product reviews Concurrency and Computation- Practice and Experience, Wiley 2018, e4956, 1-15. http://doi.org/10.1002/cpe.4956
[30] https://nid.iicm.tugraz.at/Home/BookDetail/187
[31] Mester, G. (2016). Rankings Scientists, Journals and Countries using h-Index. Interdisciplinary Description of Complex Systems, 14 (1), 1-9. http://doi.org/10.7906/indecs.14.1.1
[32] Mester, G. (2015). Merenje rezultata naučnog rada, Tehnika-Mašinstvo, Belgrade, Serbia, Vol. 64, No. 3, ISSN 0040-2176, pp. 445-454.
[33] Mester, G. (2011). Academic Ranking of World Universities 2009/2010, Ipsi Journal, Transactions on Internet Research, TIR, Belgrade, ISSN 1820 - 4503, Vol. 7, No. 1, pp. 44-47.
[34] Mester, G. (23 - 26. 02. 2015). Novi trendovi naučne metrike, Proceedings of the XXI Skup Trendovi Razvoja: "Univerzitet u Promenama...", TREND 2015, Zlatibor, Serbia, ISBN 978-86-7892-680-8, DOI: 10.13140/RG.2.1.1754.2486, paper No. UP 1-3, pp. 23-30.
[35] Namik D, Martin E, Hermann M, Bilal Z (2020) Experiences Based on a Major Information Server. IPSI BgD Transactions on Internet Research 16 (1):68-75