Topic Trends on the Hungarian Internet – News and Academic Web Presence

EXTENDED ABSTRACT

György Kampis*, László Gulyás†*, Zsolt Jurányi*

* Lorand Eötvös University, Budapest, Hungary
† AITIA International, Inc., Budapest, Hungary
E-mail: gulya@hps.elte.hu,

Abstract—Online media venues change rapidly. New content and topics emerge and disappear as the joint interest of the community producing and consuming them changes. Archiving online content is generally not solved, so the dynamics of this public discourse is rarely studied. While archiving Internet content is a vast challenge, the continuous extraction of topic categories and the archival of them could result in a trace of public interest and its dynamics over a period of time. This paper reports on a pilot study in Hungary, targeting the studying of public internet content of public news portals and of academic research institutions, and presents some early analysis results, and lessons learned. We first observe that the internet based "big data" is unexpectedly small for Hungary. Furthermore, we discuss our tools developed for the analysis and show that the dataset of the online content of the Hungarian academic institutions changes at a low rate. We suggest that differences in the productivity of the institutions can be correlated with the differences in content refreshment in their internet presence.

I. Introduction

Public discourse is changing rapidly and is often hard to analyze in retrospect. In the pre-Internet era, such projects would involve massive studies of the archives and would typically span over a long period of time. Nowadays, a vast amount of public content exists in an electronic format and is mostly available online. Therefore, the processing of this large dataset should be amenable to automation. However, the framework of public archiving was created in the era of the hard copies and it is still struggling to cope with the challenges of our days. The situation varies across countries, but no general solution exists to date (general archives such as www.archive.org/ signify a partial solution only and are complemented by various national efforts, such as www.webarchive.org.uk/ in the UK). In particular, no form of archiving of public online content currently exists in Hungary and as a partial consequence no established form of (semi-) automated analysis of the online public discourse is available for the Hungarian content.

Our goal is to develop tools and methods to collect and analyze publicly available online content in Hungary, with a special focus on the dynamics of the leading topics of the public discourse. The developed tools are applicable to general trending of the topics of online content as testified by our demonstrator study of major online news venues. We also apply our tools and methods to the online presence of Hungarian academic institutions. This latter study allows us to address questions like how the online activity of institutions corresponds to their scientific success (e.g., citation efficiency), etc.

This paper describes our basic approach and methods and provides an initial glimpse at our experiences and findings.

II. Materials and methods

We use modified harvesting techniques originally tested and tried by several earlier national archives, including the internet archiving project of the British Library. We use the Heritrix crawler modified and specially configured for our purposes. Our hardware configuration is a Dell T710 server (2x4 core Xeon E5520, 48GB RAM, 2TB HDD).

In the current study we concentrated on two major Hungarian news venues and web data obtained for 48 academic (that is, higher education and Hungarian Academy of Science, or HAS) institutions. The lists of these are: http://mta.hu/mta_kutatointezetei http://hu.wikipedia.org/wiki/Magyarországi_egyetemek_listája, and http://hu.wikipedia.org/wiki/Magyarországi_főiskolák_listája respectively.

The list of all these academic institutions is conveniently summarized at http://www.hungarianscience.org.

Files downloaded are mainly text files and videos stored at the above sites; in particular all files with the extensions exe, gz, iso, jar, mp3, ogg, ppt, rar, wav, xls, xlsx, and zip are exluded (as a reponse to the existence of many shared disk images and other large files of dubious origin that are often unrelated to the „official" activity of the downloaded sites).

An overview picture of the Hungarian Internet backbone serving the sampled institutions is shown on Figure 1.

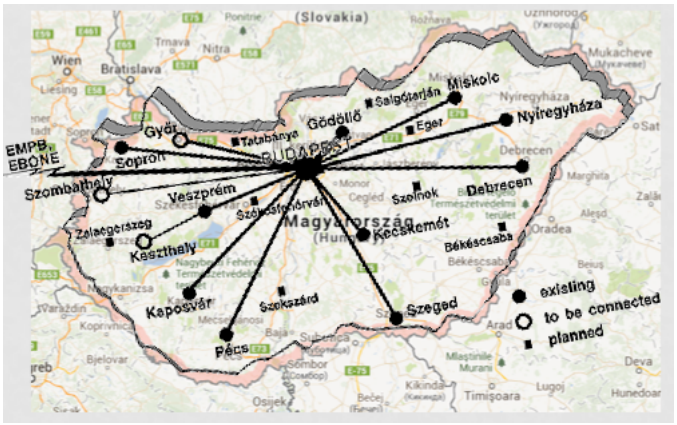


Fig. 1. The Hungarian backbone connecting cities of primary academic importance.

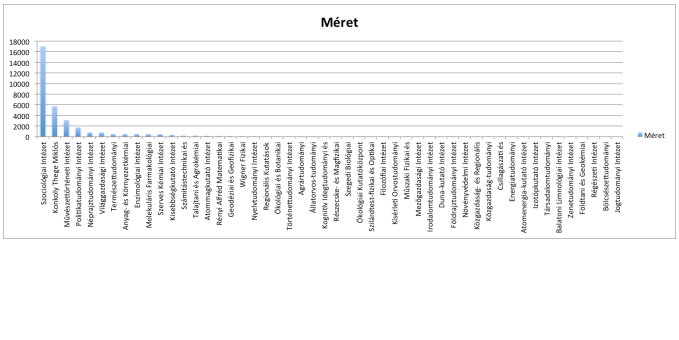
III. Results

In this paper we present our tools and statistics and demonstrate them using data from the online news venues. This will be followed by a more detailed discussion of our study of the entire record of all HAS institutions.

A preliminary analysis of the data shows that the dataset is 33GB, of this in various text formats (html, doc, docx, rtf, pdf, ps) 6,5GB. A complete copy of the entire record of all higher education institutions is 53GB, of this, text is 36GB.

A. Rank distribution of the complete records

The rank distribution of the data sizes of a snapshot of online content for the academic institutions follows a „power-law-like" distribution, as can be expected (Fig.2.)



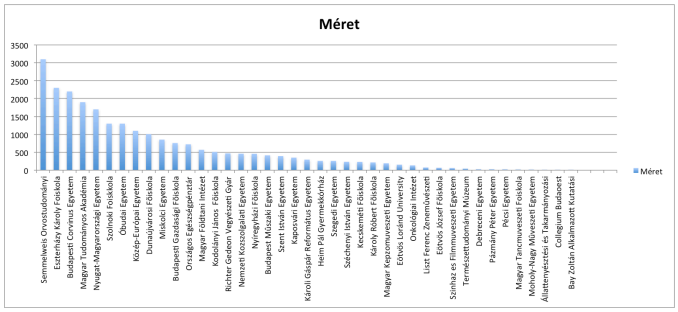


Fig. 2. Rank distributions of a snapshot of HAS (top) and higher education institutions (bottom).

B. Rank distribution of text files

A similar picture is obtained if only text files are considered (Fig.3.)

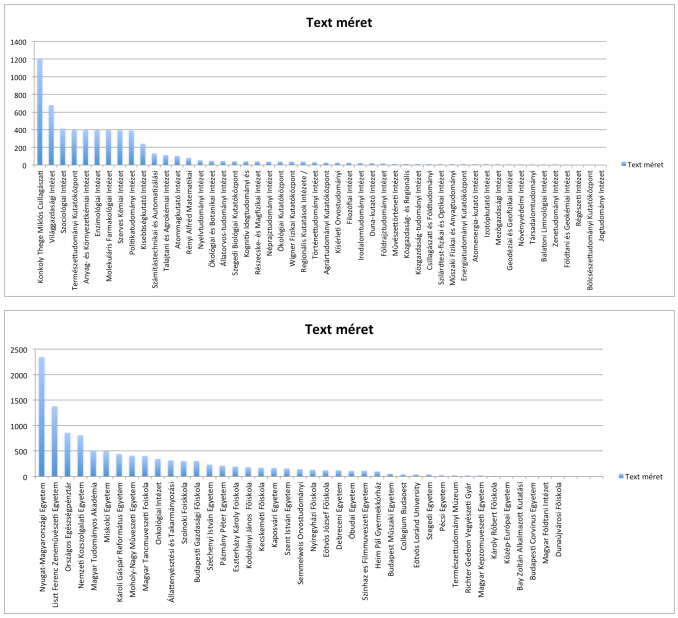


Fig. 3. Rank distributions of the text records of HAS (top) and higher education institutions (bottom).

C. Average sizes of the academic dataset

Translating the above to numbers, and concentrating on the averages we find these:

- Average size: 974 MB per site (median: 137 MB)
- Average text size: 474 MB per site (median: 47 MB)

It is well known that averages of skewed distributions should be treated with special caution. Yet the average is a safe upper estimate to the intuitive notion of "centre" and the median is often too low in these cases to be of practical use.

By comparison to these numbers, it should be noted that the text size of the personal page of one of the authors alone is 180MB.

When contemplating these figures, then, an inevitable conclusion would be that "big data" is, for the Hungarian academic institutions currently at least, rather small. This further indicates a lack of tools, competence, interest or available content (or a combination of these) for a more massive web presence of these institutions.

D. The Topics of the Academic Sites

To come.

IV. Discussion

We have presented a report on a pilot study in Hungary, targeting the studying of public Internet content of public news portals and of academic research institutions. We presented our tools and methods and demonstrated them on datasets obtained from major Hungarian online news venues and from the online presence of Hungarian academic institutions.

Our first observation is that the Internet based "big data" is unexpectedly small for Hungary. Furthermore, we observed that the dataset of the online content of the Hungarian academic institutions changes at a low rate. We suggest that differences in the academic productivity of the institutions can be correlated with the differences in content refreshment in their Internet presence.

Acknowledgment

This work was partially supported by the European Union and the European Social Fund through project FuturICT.hu (grant no.: TAMOP-4.2.2.C-11/1/KONV- 2012-0013).

References

[1] G. Kampis, 2013: Innovation Acceleration by Public Data Analysis, presentation at the FuturICT.hu "Networking Conference", Budapest, 14th June 2013.

[2] G. Kampis, L. Gulyas and S. Soós 2012: Megjósolható-e a ráfordítás a sikerből?, előadás az V. Emergencia Workhopon, Budapest, 2012 december 7.

[3] G. Kampis 2013: Approaches to activity mapping and performace evaluation in scientific production, talk given at the University of Duisburg, Aug 8., 2013.