Knowledge Capture from Multiple Online Sources with the Extensible Web Retrieval Toolkit (eWRT)

Albert Weichselbraun
University of Applied Sciences Chur
Faculty of Information Science
Pulvermühlestr. 57, 7004 Chur, CH
+41 81 2863 727
albert.weichselbraun@htwchur.ch

Arno Scharl
MODUL University Vienna
Department of New Media Technology
Am Kahlenberg 1, 1190 Vienna, AT
+43 1 3203555 500
arno.scharl@modul.ac.at

Heinz-Peter Lang
Vienna Univ. of Economics & Business
Research Institute for Comp. Methods
Augasse 2-6, 1090 Vienna, AT
+43 1 31336 5231
heinz.lang@wu.ac.at

ABSTRACT

Knowledge capture approaches in the age of massive Web data require robust and scalable mechanisms to acquire, consolidate and pre-process large amounts of heterogeneous data, both unstructured and structured. This paper addresses this requirement by introducing the *Extensible Web Retrieval Toolkit* (eWRT), a modular Python API for retrieving social data from Web sources such as *Delicious, Flickr, Yahoo!* and *Wikipedia*. eWRT has been released as an open source library under GNU GPLv3. It includes classes for caching and data management, and provides low-level text processing capabilities including language detection, phonetic string similarity measures, and string normalization.

Categories and Subject Descriptors

H.3.4 Systems and Software: Information Networks; *H.3.5 Online Information Services:* Web-based Services; *I.7.5 Document Capture:* Document Analysis.

General Terms

Algorithms, Measurement, Human Factors, Languages.

Keywords

Data acquisition, knowledge extraction, text mining, structured and unstructured information sources, social media.

1. INTRODUCTION

The rapid growth and fragmented character of social media and publicly available structured data challenges established approaches to knowledge capture. Among the research challenges when processing the acquired content are aggregating noisy input data and generating annotations of consistent high quality even when dealing with incomplete or heterogeneous data. This requires a sequence of inter-linked processing steps including data acquisition (e.g. hierarchical or focused crawling), clean-up (e.g. duplicate detection) and filtering (e.g. domain relevance checks), followed by named entity recognition and advanced relation extraction. Such sequential processing pipelines represent the core of knowledge acquisition and language processing infrastructures such as GATE [1] and webLyzard [29]. The open-source library presented in this paper complements these frameworks by providing a standardized approach to capturing and integrating the required input data from unstructured, structured and social sources based on locality sensitive hashes [19]. This *Extensible Web Retrieval Toolkit* (eWRT) provides a modular Python API for retrieving social data from Web sources such as *Delicious* [17]*, Flickr [20], Yahoo!* [31] and *Wikipedia [30]*. The API also includes classes for effective caching and data management.

The remainder of the paper is organized as follows. Section 2 presents the features of eWRT. Section 3 summarizes the process of capturing knowledge from different Web sources and discusses problems that arise from a multi-source approach. Section 4 outlines the integration process for documents and messages from multiple sources and presents with Media Watch on Climate Change a concrete use-case. Section 5 concludes the paper and sheds light on possible future research avenues.

2. TOOLKIT FEATURES

eWRT has been jointly developed by researchers at MODUL University Vienna, the University of Applied Sciences Chur, and the Vienna University of Economics and Business. The toolkit provides modules for content acquisition and caching, low-level natural language processing such as language detection, phonetic string similarity measures, methods for string normalization and methods for other related functionalities. This section provides an overview on eWRT's main features.

Content Acquisition. The eWRT.ws package contains the content acquisition components. Currently, it supports ten different social media sources including *delicious, Facebook, Flickr*, *Twitter*, and *Wikipedia*. All components access these services through a central eWRT class (eWRT.access.http) or official API clients from the provider. Enforcing rate and bandwidth limits ensures conformance with the information provider's terms of services. Due to rate limits, eWRT provides streaming APIs where possible, since they provide new content in real time and create a much lower load on the content provider's infrastructure. eWRT distinguishes two types of content acquisition components:

- modules implementing the *WebDataSources* interface, which retrieve Web content and a rich set of metadata; e.g., document title, publishing date, abstract, copyright information, links to related resources, content ratings, etc.; and
- components based on the *TagInfoServices* interface that are limited to providing data on the distribution of query terms in Web sources. These data is then used by high-level modules that compute association metrics such as Web distance for ontology learning, ontology enrichment and refinement [14].

K-CAP '13, June 23-26, 2013, Banff, Canada.
Copyright (c) 2013 ACM 978-1-45-03-2102-0/13/06… $15.00.

Content Caching. The eWRT framework provides methods for transparently caching arbitrary function calls. Exposing this functionality over decorators in accordance to the aspect-oriented programming paradigm allows the caching of CPU, memory, and time-intensive methods with a minimum overhead. Function arguments and results are recorded in a memory or disk cache allowing eWRT to answer identical queries directly from the cache. The transparent caching framework of eWRT has proven its utility, especially in conjunction with *TagInfoService* calls that often contain similar queries to external Web services.

3. CONTENT ACQUISITION

The following section introduces the webLyzard content acquisition pipeline, and outlines the usage of eWRT in this process. The pipeline includes a portfolio of modular components for acquiring unstructured and structured content.

3.1 Unstructured Content

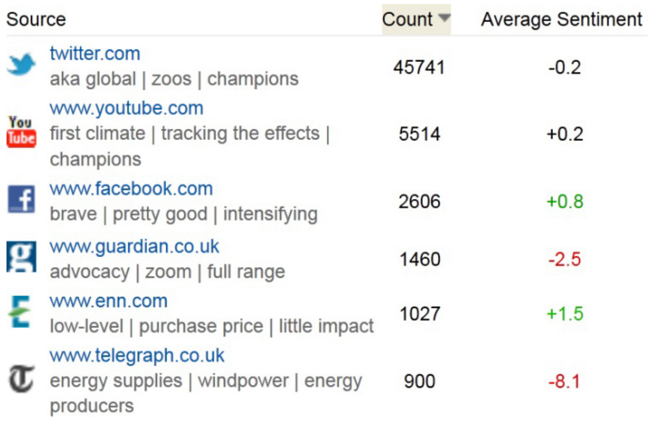
Knowledge capture from unstructured content sources involves handling massive amounts of volatile Web content and requires a scalable and flexible approach that addresses the challenges outlined below for timely acquiring and integrating Web content from heterogeneous sources.

Content Volatility. News media sites are characterized by frequent context updates; the same is true for social media sites and micro-blogging services such as Twitter. Using traditional crawling techniques on such sites is not feasible. Consequently, we prefer analyzing feed data, such as RSS feeds and the content streams published by Twitter or Facebook, since they are limited to new or updates documents only. Feeds often also contain additional metadata that is useful for the content annotation process.

Missing Content. Many sites do not publish all new content in RSS feeds, but rather a selection of what they consider the most important news items. The webLyzard content acquisition pipeline addresses this problem by (i) constantly retrieving and analyzing content feeds, and (ii) running regular mirrors (e.g. in daily or weekly intervals) that retrieve missing pages and complement articles obtained from RSS feeds.

Noise. Most Web documents contain noise that can disturb text mining methods – e.g., navigation toolbars, dynamic content elements such as headlines and related links, advertisements or copyright notices. Heuristic content extraction methods can (i) identify and extract the relevant content elements and (ii) eliminate overview pages – i.e., dynamically generated pages that merely summarize the content of other pages [4].

Outdated Content. RSS feeds often contain metadata that describes a document's creation date. HTML documents, in contrast, frequently do not publish such information, or even worse may provide incorrect date metadata. For example, many sites return either no or incorrect last-modified headers, and contain misleading date meta-tags. A heuristic date extraction component, that considers the Web page's language and region to correctly interpret ambiguous date specifications such as "01/02/2013" parses Web sites and adjusts their date, where necessary. Although the date extraction automatically handles a considerable fraction of all sites correctly (an evaluation carried out on 3187 manually labeled English and German speaking Web sites showed that the correct date was extracted from more than 78% of all sites; less than 4% yielded an invalid date and for 18% we couldn't extract a date suggestion), we still need site-specific extensions for sources that do not conform to common Web design practices.



Source	Count	Average Sentiment
twitter.com aka global | zoos | champions	45741	-0.2
www.youtube.com first climate | tracking the effects | champions	5514	+0.2
www.facebook.com brave | pretty good | intensifying	2606	+0.8
www.guardian.co.uk advocacy | zoom | full range	1460	-2.5
www.enn.com low-level | purchase price | little impact	1027	+1.5
www.telegraph.co.uk energy supplies | windpower | energy producers	900	-8.1

Figure 1. Frequency and average sentiment of "climate change" in news and social media coverage, including the three top keywords that each outlet associates with the topic (Jan-Dec 2012)

3.2 Structured Content

Providing a generic framework for gathering unstructured data from multiple sources allows enriching the captured knowledge with data from structured sources including thesauri, ontologies and Linked Data. Relevant data from structured sources such as *DBpedia* [16], *Freebase* [21] and *OpenCyc* [23] helps to enrich concepts and instances with high-quality background information identified through contextualised keyword searches and conceptual matching [5; 9; 13]. Online ontologies can be integrated through semantic web search engines such as *Sindice* [25] and *Watson* [28], while SPARQL [26] query interfaces allow accessing relevant Linked Data sources. Earlier work on *Scarlet* [36] provides algorithms for (i) collecting and disambiguating structured data; (ii) fusing structured elements referring to the same entity; (iii) associating these aggregated structures to relevant entities in the *Media Watch on Climate Change,* and (iv) detecting and reconciling knowledge structures that describe the same entities by means of cross-lingual alignment.

Developed as part of the DIVINE research project [18], the structured data acquisition component has been applied to several domains including climate change [13] and tourism [7].

4. CONTENT INTEGRATION

The methods described above yield an abundance of heterogeneous data that needs to be integrated in a consistent way. Although multi-source data acquisition provides more comprehensive coverage and shorter delays between publication and analysis, it is prone to include duplicate documents in the content repository since identical or very similar content is retrieved from different services or URLs, containing minor changes from "related content" sections or dynamic data fields, which renders standard hashing methods useless. webLyzard removes duplicates by using a Java implementation of *Nilsimsa,* an algorithm that computes 256-bit locality sensitive hash values [6]. The bitwise difference between such hashes provides an estimation of how different the respective input documents are (small changes yield similar Nilsimsa hashes) and allows identifying documents and messages that humans would perceive as identical even if they contain a certain level of noise (e.g. OCR errors) or minor editorial updates. We use highly optimized low-level operators such as binary XOR in conjunction with the *popcount* algorithm [3] to compare and integrate documents from different input streams with content already present in the webLyzard knowledge repository.

4.1 System Overview

Figure 2 outlines the data acquisition and consolidation process in more detail. The distributed webLyzard Web crawler, which is run in regular intervals, retrieves Web documents based on the sample specifications in its database. After the date extraction yields a corrected document date, the content cleanup component identifies and extracts the relevant content from the Web page. An RSS Feed Extractor dynamically scans the mirrored content for new RSS feeds that are then included in the RSS feed table. The RSS crawler continuously monitors and fetches those feeds, providing new or updates documents to the content cleanup component.

In addition to content retrieved from unstructured sources, eWRT components obtain a massive amount of documents and messages from social media sources. A merging component integrates the content from all input streams (Web crawler, RSS Feed Crawler, and multiple eWRT input streams) and removes duplicate messages, before the content is forwarded to the natural language processing (NLP) pipeline. The subsequent NLP and text mining components draw upon the data obtained from structured sources to improve low level content annotation and text extraction processes – e.g. using context information from *GeoNames* to improve and evaluate the geo-tagging component [10], or leveraging *ConceptNet, DBpedia, SenticNet* and *SentiWordnet* for ontology learning [13] and concept-aware sentiment analysis [2; 12].

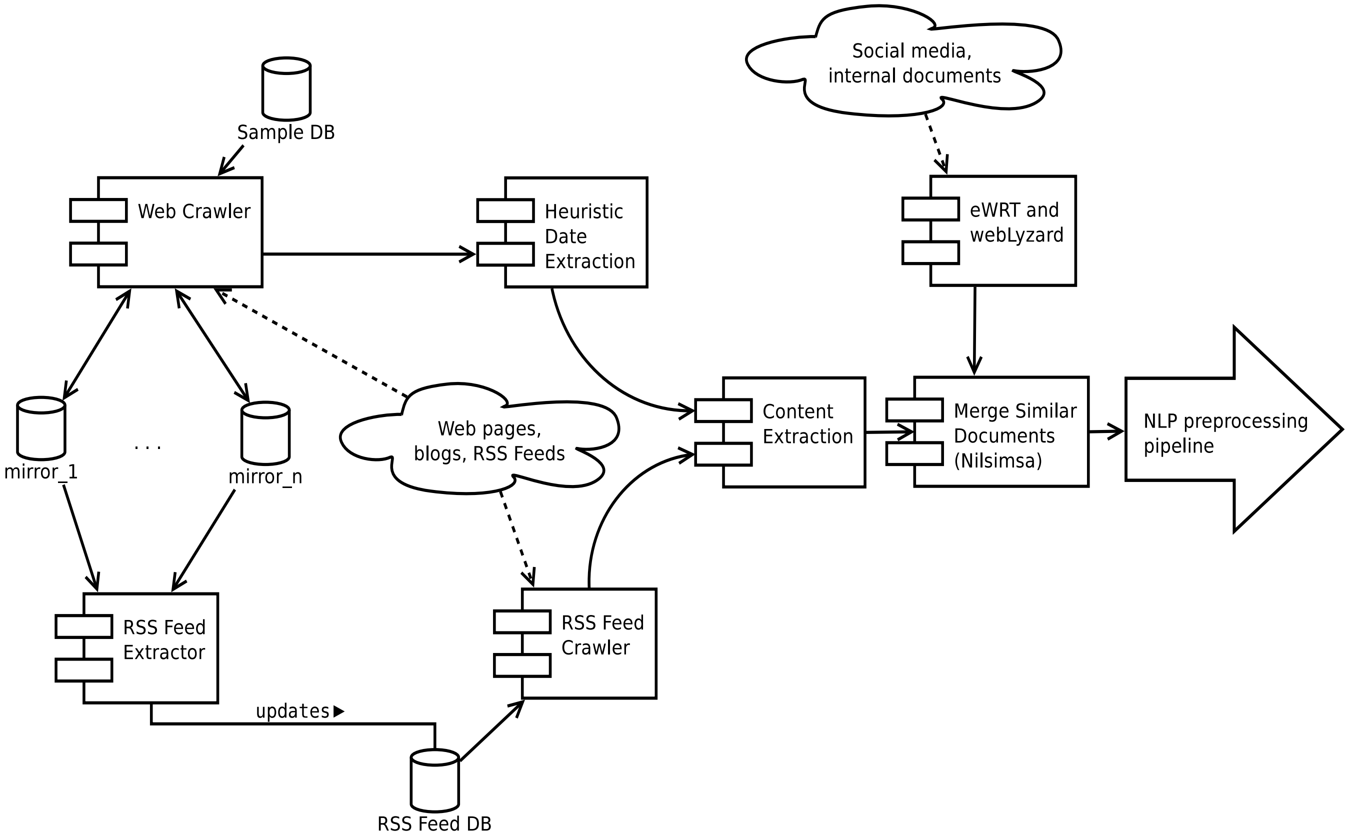


Figure 2. Integrating RSS feeds with crawled Web content to improve the knowledge extraction process

4.2 Climate Change Use Case

The *Media Watch on Climate Change* [8; 22] is a public Web portal that investigates online coverage on climate change and related environmental issues (see Figure 1). The portal uses eWRT in conjunction with streaming APIs to capture various types of Web content including news and social media, environmental NGOs, and the Web sites of Fortune 1000 companies.

The required sentence-level and document-level metadata is generated by the annotation mechanisms of the webLyzard Web intelligence platform [29]. These mechanisms continuously update and structure the climate change-related content repository. The gathered content, about one million documents per week, is managed with the *PostgreSQL* [24] relational DBMS and indexed using *Apache Lucene* [15].

5. CONCLUSION AND OUTLOOK

Future research will use document-level metadata to guide the distributed data acquisition process, exploiting commonalities of metadata elements to group information objects from different evidence sources along multiple semantic dimensions such as time, location and prevalent topic. This not only allows determining the domain relevance of a document computationally (and thus decide upon its inclusion into the repository of the Media Watch on Climate Change), but also to exploit the potential of caching and resource allocation mechanisms in future eWRT releases. These mechanisms will be based on the *Search-Test-Stop (STS)* model [11] to address computational bottlenecks and optimize queries according to their specific costs in terms of request time and resource usage.

The *uComp Project* [27] builds upon the emerging field of *Human Computation* to increase the scalability of existing knowledge capture and extraction approaches, gathering collective intelligence from large user communities. uComp will put considerable effort in extending not only the functionality, but also the documentation of eWRT in order to promote the adoption and further development of the framework.

6. ACKNOWLEDGEMENT

The research presented in this paper has been conducted as part of the DIVINE Project (www.weblyzard.com/divine) funded by the Austrian Research Promotion Agency, the WISDOM Project (www.semanticlab.net/wisdom) funded by the Swiss Commission for Technology and Innovation, and the uComp Project (www.ucomp.eu) funded by the Austrian Science Fund within the European CHIST-ERA program line.

7. REFERENCES

[1] Cunningham, H., Tablan, V., Roberts, A. and Bontcheva, K. (2013). “Getting More Out of Biomedical Documents with GATE's Full Lifecycle Open Source Text Analytics”, *PLoS Computational Biology*, 9(2): 1-16.

[2] Das, A. and Gambäck, B. (2012). Sentimantics: Conceptual Spaces for Lexical Sentiment Polarity Representation with Contextuality. *3rd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA-2012)* Jeju Island, Korea: Association for Computational Linguistics: 38-46.

[3] Knuth, D.E. (2009). *The Art of Computer Programming (Volume 4, Fascicle 1: Bitwise Tricks & Techniques; Binary Decision Diagrams)*. Reading: Addison-Wesley Professional.

[4] Lang, H.-P., Wohlgenannt, G. and Weichselbraun, A. (2012). TextSweeper - A System for Content Extraction and Overview Page Detection. *International Conference on Information Resources Management (Conf-IRM 2012)*. Vienna, Austria: Association for Information Systems.

[5] Mendes, P.N., Jakob, M., García-Silva, A. and Bizer, C. (2011). DBpedia Spotlight: Shedding Light on the Web of Documents. *7th International Conference on Semantic Systems (I-Semantics 2011)*. Graz, Austria: 1-8.

[6] Paulevé, L., Jégou, H. and Amsaleg, L. (2010). “Locality Sensitive Hashing: A Comparison of Hash Function Types and Querying Mechanisms”, *Pattern Recognition Letters,* 31(11): 1348-1358.

[7] Sabou, M., Arsal, I. and Brasoveanu, A.M.P. (2013). “TourMISLOD: A Tourism Linked Data Set”, *Semantic Web Journal*: Forthcoming.

[8] Scharl, A., Hubmann-Haidvogel, A., et al. (2013). Media Watch on Climate Change – Visual Analytics for Aggregating and Managing Environmental Knowledge from Online Sources. *46th Hawaii International Conference on Systems Sciences (HICSS-46)*. R.H. Sprague. Maui, USA: IEEE Press: 955-964.

[9] Waitelonis, J., Ludwig, N., Knuth, M. and Sack, H. (2011). “WhoKnows? - Evaluating Linked Data Heuristics with a Quiz that Cleans Up DBpedia”, *International Journal of Interactive Technology and Smart Education,* 8(4): 236-248.

[10] Weichselbraun, A. (2009). A Utility Centered Approach for Evaluating and Optimizing Geo-Tagging. *1st International Conference on Knowledge Discovery and Information Retrieval (KDIR-2009)*. Madeira, Portugal: 134-139.

[11] Weichselbraun, A. (2010). “Optimizing Queries to Remote Resources”, *Journal of Intelligent Information Systems,* 37(2): 119-137.

[12] Weichselbraun, A., Gindl, S. and Scharl, A. (2013). “Extracting and Grounding Context-Aware Sentiment Lexicons”, *IEEE Intelligent Systems*: Forthcoming (Accepted 06 Jan 2013).

[13] Weichselbraun, A., Wohlgenannt, G. and Scharl, A. (2010). “Refining Non-Taxonomic Relation Labels with External Structured Data to Support Ontology Learning”, *Data & Knowledge Engineering,* 69(8): 763-778.

[14] Weichselbraun, A., Wohlgenannt, G. and Scharl, A. (2011). Applying Optimal Stopping Theory to Improve the Performance of Ontology Refinement Methods. *44th Hawaii International Conference on System Sciences (HICSS-44)*. Kauai, USA.

Online Resources

[15] Apache Lucene. http://lucene.apache.org/.

[16] DBpedia. http://www.dbpedia.org/.

[17] Delicious. http://www.delicious.com/.

[18] DIVINE Research Project. http://www.weblyzard.com/divine/.

[19] Extensible Web Retrieval Toolkit (eWRT). http://www.weblyzard.com/ewrt/.

[20] Flickr. http://www.flickr.com/.

[21] Freebase. http://www.freebase.com/.

[22] Media Watch on Climate Change. http://www.ecoresearch.net/climate/.

[23] OpenCyc. http://www.opencyc.org/.

[24] PostgreSQL. http://www.postgresql.org/.

[25] Sindice. http://www.sindice.com/.

[26] SPARQL Inferencing Notation (SPIN). http://www.spinrdf.org/.

[27] uComp Research Project. http://www.ucomp.eu/.

[28] Watson. http://watson.kmi.open.ac.uk/WatsonWUI/.

[29] webLyzard. http://www.webLyzard.com/.

[30] Wikipedia. http://www.wikipedia.org/.

[31] Yahoo! http://www.yahoo.com/.