Shortipedia
Aggregating and Curating Semantic Web Data

Denny Vrandečić[a], Varun Ratnakar[b], Markus Krötzsch[c], Yolanda Gil[b]

[a] *Institute AIFB, KIT Karlsruhe Institute of Technology, Karlsruhe, Germany,* denny.vrandecic@kit.edu
[b] *ISI, USC University of Southern California, Marina del Rey, CA, USA,* {varunr|gil}@isi.edu
[c] *Computing Laboratory, University of Oxford, Oxford, UK,* markus.kroetzsch@comlab.ox.ac.uk

Abstract

Shortipedia is a Web-based knowledge repository, that pulls together a growing number of sources in order to provide a comprehensive, diversified view on entities of interest. Contributors to Shortipedia can easily add claims to the knowledge base, provide sources for their claims, and find links to knowledge already available on the Semantic Web.

Keywords: semantic wikis, data integration, semantic infrastructure technology, technology for collaboration

1. Introduction

The goal of Shortipedia is to develop a reference Web site that combines automatic aggregation of facts on the Semantic Web and manual curation by volunteer users. Since different sources can take different views and contain inconsistent information, central to the design of Shortipedia is that any assertion is in principle possibly true and must be included in the repository together with its provenance. Instead of taking assertions as true factual knowledge, we interpret them as assertions about an entity that are true within the context of their provenance record.

Key features of Shortipedia are that it

1. assigns a unique identifier to entities of interest using DBpedia [4] as the basis for identity and Wikipedia as the basis for consensus on entities of interest,
2. collects those identifiers in a consensus-built resource,
3. aggregates facts about those entities from the Web of Data and other available data sources,
4. embraces the diversity of views on any given fact and documents their provenance, and
5. supports the curation of its contents by volunteers.

Shortipedia is built on top of a number of Web services and frameworks in order to provide a comprehensive and rewarding workflow for contributors to the project, and thus ultimately to create a reference website for all kind of facts. Whereas DBpedia already shows the high potential of providing a nucleus for the Web of Data, an *editable* extension of such a nucleus could join the advantages of Wikipedia and DBpedia, building on both services and enriching them.

As a knowledge repository, Shortipedia contains the following types of information:

- a core of *entities*, based on Wikipedia articles,
- *mappings* of those entities to entities in the Web of Data,
- aggregated *assertions* from Web of Data sources,
- *provenance records* of all assertions, and
- direct *contributions* and corrections from users.

While the first four types of information deal with aggregating and curating Semantic Web data, the last type – the direct contributions – allow users the immediate addition and correction of knowledge to Shortipedia, thus providing a Wikipedia-like incentive and growth system. This turns Shortipedia into a system for knowledge acquisition, where the provenance of the assertion is provided by the user editing log.

This results in a system that provides

- a knowledge repository containing an aggregated and curated view on Semantic Web data,
- a query and browsing interface for Linked Open Data, and
- a neutral and unbiased reference site that includes alternative points of view.

Shortipedia was publicly released as a beta in November 2010 and the experiences gathered with Shortipedia currently inform the development of a similar project based on a more scalable approach. This system description first gives an overview of the architecture and some related approaches in Section 2 and 3. In Section 4 we note some of the design decisions we have made, followed by a summary of lessons learned in Section 5. The system is available online at `http://shortipedia.org`.

2. Architecture

Shortipedia is an application of MediaWiki, a highly extensible, open source wiki software [2]. MediaWiki is widely used both on the Web, most famously on Wikipedia [1], as well as in corporate intranets for a huge variety of tasks. In order to support these tasks, more than 1,200 extensions are being provided for download from the MediaWiki homepage.[1] Shortipedia is based on a newly developed extension that can be deployed on top of existing MediaWiki installations. The system builds upon functionality that is provided by a number of existing MediaWiki extensions. We will describe both the newly developed software and its relationship to existing components in this section.

The major extension used in Shortipedia is Semantic MediaWiki (SMW). SMW enables adding metadata to wiki pages and querying the collected metadata of the wiki [15]. Due to this functionality, the wiki can also be regarded as a light-weight, schemaless database editable through the wiki user interface. So in addition to the richly formatted, human-readable wiki pages about persons, cities, companies, music bands, etc., the system allows machine-readable metadata to be associated with these pages. The particular strength of SMW is that it closely integrates both aspects, enabling the collaborative maintenance of semantic data as part of the established editing process. Various querying and browsing features provide immediate means for re-using the data within the wiki – a crucial incentive for editors to contribute metadata in the first place. Further details can be found in [15]. Shortipedia builds on the technical infrastructure of SMW but replaces the normal text-based editing interface of MediaWiki (and thus Semantic MediaWiki) with a purely form-based interface. The template syntax of MediaWiki is heavily used to create a wiki that is primarily aimed at editing data instead of text.

Building on this baseline, we already have a system that is known to be scalable with regards to the size of the data and the size of the community it is supposed to handle, and that provides us with all the basic functionalities like user management, rights management, an infrastructure for Restful APIs, complete and rich article history features, and a comprehensive extension framework.

Moreover, SMW is part of a vibrant ecosystem of related extensions, that can be used to further extend the services provided by Shortipedia. The functionality provided by available extensions ranges from additional query services, such as the rich query capabilities of *Ask the Wiki* [13] or Halo[2], to support for cross-site data exchange and synchronization, as provided by *DataPress*[3] or *Distributed SMW* [20]. Most of these extensions can be readily deployed in our scenario, but we focus on the core functionality of Shortipedia within this system description.

For Shortipedia we developed a MediaWiki extension to provide us with the missing features we need for the system. These include the following:

Loading Data from the Semantic Web.
Information that is available as Linked Open Data [3] is integrated into the system. This data is then visualized for the end user, and can be mapped to the wiki-internal vocabulary. The information can then be copied into the wiki system while retaining the link to the original source as a reference.

Simple Editing of Facts and References.
Shortipedia features a custom user interface, so that no wiki syntax needs to be learned by the user. The new interface completely

[1] `http://www.mediawiki.org/wiki/Manual:Extensions`

[2] `http://smwforum.ontoprise.com/`

[3] `http://projects.csail.mit.edu/datapress/`

Figure 1: The English Shortipedia page for Shanghai `http://shortipedia.org/topic/en/Shanghai`, including Wikipedia article text (left), facts gathered by Shortipedia (top right), and related entities on the Web of Data (bottom right)

replaces the classical editing of the wiki's markup language, and the latter is disabled.

Multi-language support. Support for internationalization generally increases usability, but is of particular relevance for establishing an integrated knowledge base like Shortipedia.

Each entity in Shortipedia is represented by a dedicated page, as illustrated in Fig. 1. If available, the page displays the related Wikipedia article for the entity. Wikipedia and DBpedia [4] together have created the largest resource of community-built consensus on the meaning of Web identifiers. For bootstrapping, we assume the mapping that DBpedia provides for Wikipedia articles, i.e. we assume that the Wikipedia article at `http://en.wikipedia.org/wiki/X` is about the entity referenced by `http://dbpedia.org/resource/X`.

Using the DBpedia identifier, we then query the *sameAs.org* service [12]. The latter is a service that, given an URI, provides further URIs that are considered co-referent. Moreover, the Semantic Web search engine *Sindice* provides entities given a search term, among other services [18]. We thus also query Sindice for data about the title of the entity (the Wikipedia page name). Shortipedia is designed to allow further sources of identity to be incorporated later, e.g., for special fields such as protein research (e.g. through Chem2Bio2RDF [7]) or music (e.g., MusicBrainz [22]).

To process the data that is retrieved from the Semantic Web, Shortipedia incorporates the RDF parser ARC2.[4] This allows us to dereference the URIs suggested by sameAs.org and Sindice, and to provide retrieved data to the user upon request. An example is shown in Fig. 1, where the entity retrieved from New York Times (bottom right) has been expanded. We use AJAX (via the jQuery library[5]) to execute such requests dynamically. This requires the information about the respective entities to be available as Linked Open Data. The system then displays the assertions gathered from the Web of Data, and allows the user to map external vocabulary to Shortipedia vocabulary. The

[4] `http://arc.semsol.org/`
[5] `http://jquery.com/`

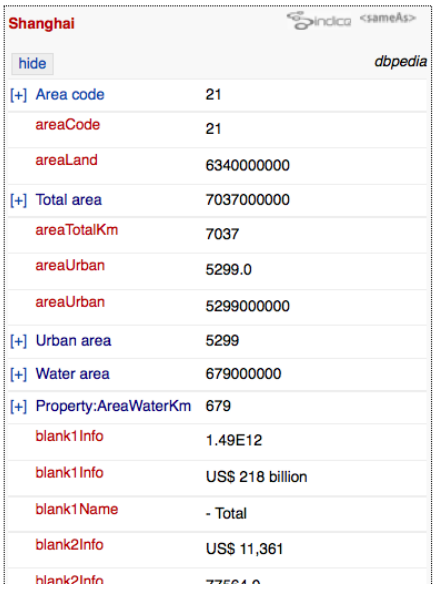

Figure 2: View on DBpedia data on the English Shortipedia page for Shanghai

respective interface is shown in Fig. 2. Blue terms have been mapped to the internal vocabulary, while red terms still need to be mapped. Mappings are managed globally, so that every external vocabulary URI needs to be mapped only once. When the vocabulary of a fact has been mapped, the user can add it to the Shortipedia knowledge base by clicking on the plus icon in Fig. 2. The provenance trail of the fact is preserved in this operation.

3. Related Approaches

Various approaches of collecting structured knowledge from volunteers have been studied in prior research. The OpenMind[6] constellation of projects has been collecting common sense knowledge[7] to create structured repositories.[8] The Cyc FACTory [17] allows contributors to add facts to the Cyc knowledge base [16], but the contributions have to conform to the pre-defined schema. The ESP game[9] can also be considered to structure contributions, as it motivates users to provide common labels to pictures through rewards. Our own work on Learner [8, 9, 10] studied the contribution by volunteers of common knowledge, and reported on issues of quality of the content and the design of the user interaction to maximize the breadth of contributions as well as the automated reasoning they enable.

Sig.ma [24] provides an interface to the Web of Data, in which each user can create a live collection of data, pulling it together from different sources that can be republished in several different ways. Sig.ma does not provide a place for the collaborative curation and collection of data, but gives this power only to the individual user for the creation of their own view on the Web of Data.

There are several efforts in extracting structured knowledge from Wikipedia. The Intelligence in Wikipedia project [26, 27, 28, 29, 14] combines automated text extraction with community created content. Researchers have used structured information from Wikipedia to study relatedness [19], expertise [21], and document indexing [23].

DBpedia [4] is a project to extract structured information from Wikipedia. It provides the Web of Data with a much needed nucleus of stable, Wikipedia-based URIs. DBpedia uses a multitude of automatic and semi-automatic techniques to extract information from Wikipedia articles. Whereas Shortipedia also takes Wikipedia as its starting point for entities, especially for establishing identity, it is not based on an automatic extraction from Wikipedia, but rather on the completely human effort for extracting and curating single facts. Also, DBpedia does not provide the direct editing of its content but relies on respective changes in the original source, i.e. Wikipedia, or the extraction mechanism. Recently, Ultrapedia[10] has been released to provide a wiki-based interface on top of DBpedia data, including the possibility to easily and selectively edit the source Wikipedia text and thus provide a feedback loop to the original source.

Freebase is a Web-based database that enables the creation and editing of data entries for any entity of general interest [6]. Whereas close in spirit, Shortipedia puts a strong emphasis on retaining references for every piece of information. Freebase in turn provides a more polished user interface, but the user interface (not the underlying model) is, as of writing, only available in English, whereas Shortipedia was multilingual from the beginning. The main difference between Shortipedia and Freebase

[6] http://www.openmind.org
[7] http://openmind.media.mit.edu
[8] e.g., http://csc.media.mit.edu/conceptnet
[9] http://www.espgame.org
[10] http://wiking.vulcan.com/up/index.php/Main_Page

is the completely schemaless environment of Shortipedia, a trait inherited from Semantic MediaWiki. Freebase on the other hand uses types and their schemas for entities. This allows for a more supportive interface by Freebase, but at the same time it constraints the usage of arbitrary properties that is possible in Shortipedia. This is expected to lead to a more active community with regards to the definition of new properties in Shortipedia, but it is also expected that Freebase will in general have more complete and consistent data. Due to the limited release of Shortipedia this hypothesis cannot be tested yet.

Not surprisingly, both Freebase and DBpedia are frequent sources suggested for mappings in Shortipedia, and thus data from both efforts are reused inside the project.

Also, using external data in Semantic MediaWiki is not a new idea [11]. Instead of allowing the tight integration for querying that many of these other solutions provide, Shortipedia rather aims at further integrating external data within the system, and then to allow to query this integrated data set. Indeed, Shortipedia is rather different from typical uses of SMW, regarding both the type of information that is managed, and the associated workflow and interface. In particular, the active discovery of new data sources, and the integration of from these sources distinguishes our approach from more static cross-site data retrieval methods proposed for SMW before.

4. Design Decisions

In this section we describe some key decisions with regards to the design of Shortipedia, and our rationales behind them. We acknowledge that these decisions are not imperative, but could often been taken differently, resulting in a different system.

Wikipedia and DBpedia for identity. In order to circumnavigate the core problem of how to establish identity on the Semantic Web, we build up on the work already provided by these two established sources.

Consistency is not required. The system is designed deliberately so that it allows the contributors to add data that leads to inconsistencies. Instead of ensuring data quality by enforcing some form of logical consistency or formal constraints, we merely require data sources to be referenced well. We thus exploit the wiki paradigm of community-based editing, which shows that high quality content can be obtained by collaboratively improving individual contributions in an incremental fashion. Although we generally assume these principles to be effective for managing data as well, we expect that communities will develop "gardening" approaches that are specific to the type of content in Shortipedia. For example, possible errors can be detected more easily in machine-readable data using consistency rules, even if the system cannot repair potential problems automatically [25].

Triples are immutable. We do not allow for facts to be changed. Instead, contributors can add new facts and provide better sources, or they can delete facts. We retain a history of all edits to enable easy recovery from vandalism.

Don't trust the triples. Every fact can be given a reference, and the user can decide which referenced sources to trust. We considered implementing a rating system for triples directly, e.g., users would be able to agree/disagree/like the fact that Paris is the capital of France, etc., but we dropped that idea, since it would be unclear what such an endorsement would even mean (see also the discussion on Wikipedia and democracy [5]). Instead we focus on external references.

sameAs is different. Stating co-reference for two URIs is not regarded as merely another assertion in the system, but rather is completely materialized – meaning that, once an assertion is added to the system it does not keep track of the original URIs used to describe the assertion. Deleting the mapping to the external URI will thus not delete the assertions added while the URI was mapped. We do not expect this to become a problem since (1) facts still need to be added manually even though two URIs have been mapped, and (2) a specific external source often uses a specific URI for an entity, and since the source information is retained the URI can then be reconstructed. In order to enable this, the system also enforces that the mapping property is inverse functional, i.e. that the same external

URI cannot be mapped to two different URIs in the Shortipedia namespace.

First map, then add. Facts from the Web of Data can only be added to the system after a mapping to internal entities has been established. Both the property and the value (if not a literal) have to be mapped to URIs internal to Shortipedia. This ensures that data integration is part of the editing process, preventing Shortipedia from becoming a mere data aggregator for facts discovered on the Web.

The browser does the work. In order to provide a performant and responsive interface we delegate much of the processing to the browser. We think that a compelling UI is important to engage a broader community. The downside is that complex client-side processing may exclude some browsers or devices. To address this, we plan to add a simple browse-only interface that cannot be used to edit Shortipedia.

5. Lessons Learnt

This section discusses a number of sometimes unexpected problems that we encountered in the development of Shortipedia.

Data is noisy or hard to understand. The recent explosive growth of the Linked Open Data cloud gave us hope to find huge amounts of interesting and useful data on the Web – and indeed, huge amounts of data we found. But it is often hard to give a meaningful interpretation to this data, and some datasets, especially automatically extracted ones, sport an unfavorable amount of noise. These challenges point to interesting future work.

What's in a name? Many RDF resources on the Web only carry few or distributed labeling annotations, raising the question of how to display those entities to the end user. Whereas entities, when dereferenced, do often have some labeling annotation for themselves, they do not carry such annotations for the other entities mentioned in the RDF document. A possible solution is to dereference all mentioned entities, giving rise to the next issue.

The Semantic Web is slow. Dereferencing big numbers of entities, e.g. loading several hundred RDF documents just to create a single page, would be unacceptably slow. We thus had to find a viable balance between the number of calls and the additional information we expected from the calls. Another approach was to execute some calls only when explicitly requested by the user instead of providing all relevant data as part of the page that is initially sent to the browser.

Semantic Web browsers cause problems. We wanted to enable users to browse the referenced data sources, but faced two problems: (1) there are many different Semantic Web browsers, and linking to all of them clutters the interface, and (2) often the browsers did not work for extended periods of time, or with the given data source, or had some other problems. In order to circumvent this problem, we created a web service called the *Linked Open Data Browser Switch*,[11] a site that lets the user decide on the browser to use and remembers that decision for the next time.

6. Conclusions

Shortipedia aims to show that (1) it is possible to collect encyclopedic-style structured knowledge in the form of object-property-value triples that can be aggregated to answer structured queries, and (2) that volunteer contributors can be used to integrate and validate a variety of existing sources of such triples. It builds on the widely used Semantic MediaWiki framework, which supports the entry of structured facts about the topic of any given page. We expect Shortipedia to provide valuable feedback and experiences for the development of scalable and open collaboratively-built knowledge bases of the future.

Acknowledgements

Work presented in this paper has been funded by the EU IST FP7 project RENDER and the National Science Foundation (NSF) under Grant number IIS-0948429. We acknowledge (and appreciate) the MediaWiki and Semantic MediaWiki community for their continuous support for the SMW platform and being the incubator for many of the ideas presented in this work.

[11] `http://browse.semanticweb.org`

References

[1] Phoebe Ayers, Charles Matthews, and Ben Yates. *How Wikipedia works*. No Starch Press, San Francisco, CA, October 2008.

[2] Daniel J. Barret. *MediaWiki*. O'Reilly, 2008.

[3] Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked Data – the story so far. *International Journal on Semantic Web and Information Systems*, 2009. Special Issue on Linked Data.

[4] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia – a crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165, 2009.

[5] Laura Black, Ted Welser, Jocely DeGroot, and Daniel Cosley. "Wikipedia is not a democracy": Deliberation and policy-making in an online community. In *Annual meeting of the International Communication Association*, Montreal, Quebec, Canada, May 2008.

[6] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *SIGMOD'08: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pages 1247–1250, New York, NY, USA, 2008. ACM.

[7] Bin Chen, Ying Ding, Huijun Wang, D.J. Wild, Xiao Dong, Yuyin Sun, Qian Zhu, and M. Sankaranarayanan. Chem2bio2rdf: A linked open data portal for systems chemical biology. In Nick Cercone and Jimmy Huang, editors, *Proceedings of the International IEEE/WIC/ACM Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, Toronto, Canada, August 2010.

[8] Tim Chklovski. Designing interfaces for guided collection of knowledge about everyday objects from volunteers. In *Proceedings of the 2005 International Conference on Intelligent User Interfaces (IUI'05)*, San Diego, CA, January 2005.

[9] Tim Chklovski and Yolanda Gil. An analysis of knowledge collected from volunteer contributors. In *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI'05)*, Pittsburgh, PA, July 2005.

[10] Tim Chklovski and Yolanda Gil. Towards managing knowledge collection from volunteer contributors. In *Proceedings of the 2005 AAAI Spring Symposium on Knowledge Collection from Volunteer Contributors (KCVC)*, Stanford, CA, March 2005.

[11] Basil Ell. Integration of external data in semantic wikis. Master's thesis, Hochschule Mannheim, 2009.

[12] Hugh Glaser, Afraz Jaffri, and Ian Millard. Managing co-reference on the semantic web. In *Linked Data on the Web (LDOW'09)*, Madrid, Spain, April 2009.

[13] Peter Haase, Daniel M. Herzig, Mark Musen, and Duc Thanh Tran. Semantic wiki search. In *6th Annual European Semantic Web Conference (ESWC'09), Heraklion, Crete, Greece*, volume 5554 of *LNCS*, pages 445–460. Springer Verlag, Juni 2009.

[14] Raphael Hoffmann, Saleema Amershi, Kayur Patel, Fei Wu, James Fogarty, and Daniel S. Weld. Amplifying community content creation with mixed initiative information extraction. In *Proceedings of the 27th International Conference on Human Factors in Computing System (CHI'09)*, Boston, MA, April 2009.

[15] Markus Krötzsch, Denny Vrandečić, Max Völkel, Heiko Haller, and Rudi Studer. Semantic wikipedia. *Journal of Web Semantics*, 5:251–261, September 2007.

[16] Douglas B. Lenat and R.V. Guha. *Building Large Knowledge-Based Systems: Representation and Inference in the Cyc Project*. Addison-Wesley Publishing Company, Reading, Massachusetts, 1989.

[17] Cynthia Matuszek, Michael J. Witbrock, Robert C. Kahlert, John Cabral, David Schneider, Purvesh Shah, and Douglas B. Lenat. Searching for common sense: Populating Cyc from the Web. In *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI'05)*, Pittsburgh, PA, July 2005.

[18] Eyal Oren, Renaud Delbru, Michele Catasta, Richard Cyganiak, Holger Stenzhorn, and Giovannia Tummarello. Sindice.com: A document-oriented lookup index for open linked data. *International Journal of Metadata, Semantics and Ontologies*, 3(1), 2008.

[19] Simone Paolo Ponzetto and Michael Strube. Knowledge derived from Wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, 30(1), 2007.

[20] Charbel Rahhal, Hala Skaf-Molli, Pascal Molli, and Stéphane Weiss. Multi-synchronous collaborative semantic wikis. In Gottfried Vossen, Darrell D. E. Long, and Jeffrey Xu Yu, editors, *Proceedings of the International Conference on Web Information Systems Engineering (Wise'09)*, volume 5802 of *LNCS*, Poznan, Poland, October 2009.

[21] Terrell Russell, Bongwon Suh, and Ed Chi. A comparison of generated Wikipedia profiles using social labeling and automatic keyword extraction. In *Fourth International Conference on Weblogs and Social Media (ICWSM'10)*, Washington, DC, May 2010.

[22] Aaaron Swartz. MusicBrainz: a semantic web service. *IEEE Intelligent Systems*, 17(1):76–77, 2002.

[23] Zareen Syed, Tim Finin, and Anupam Joshi. Wikipedia as an ontology for describing documents. In *Second International Conference on Weblogs and Social Media (ICWSM'08)*, Seattle, WA, March 2008.

[24] Giovanni Tummarello, Richard Cyganiak, Michele Catasta, Szymon Danielczyk, Renaud Delbru, and Stefan Decker. Sig.ma: live views on the web of data. *Journal of Web Semantics*, 2010.

[25] Denny Vrandečić. *Ontology Evaluation*. PhD thesis, KIT Karlsruhe Institute of Technology, Karlsruhe, Germany, 2010.

[26] Daniel S. Weld, Fei Wu, Eytan Adar, Saleema Amershi, James Fogarty, Raphael Hoffmann, Kayur Patel, and Michael Skinner. Intelligence in Wikipedia. In *Proceedings of the 23rd National Conference on Artificial intelligence (AAAI'08)*, Chicago, IL, July 2008.

[27] Fei Wu, Raphael Hoffmann, and Daniel S. Weld. Information extraction from Wikipedia: moving down the long tail. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge discovery and Data Mining*, Las Vegas, NV, August 2008.

[28] Fei Wu and Daniel S. Weld. Autonomously semantifying Wikipedia. In *Proceeding of the ACM Sixteenth Conference on Information and Knowledge Management (CIKM'07)*, Lisbon, Portugal, November 2007.

[29] Fei Wu and Daniel S. Weld. Automatically refining the Wikipedia infobox ontology. In *Proceedings of the 17th International Conference on World Wide Web (WWW'08)*, Beijing, China, April 2008.