WISDOM2013: A Large-scale Web Information Analysis System

Masahiro Tanaka Stijn De Saeger* Kiyonori Ohtake Chikara Hashimoto
Makoto Hijiya Hideaki Fujii Kentaro Torisawa
National Institute of Information and Communications Technology (NICT)
3-5 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 219-0289, Japan
{mtnk, stijn, kiyonori.ohtake, ch, hijiya, h-fujii, torisawa}@nict.go.jp

Abstract

We demonstrate our large-scale web information analysis system called WISDOM2013, which consists of several deep semantic analysis systems such as a factoid QA, a non-factoid QA and a sentiment analyzer, and a software platform on which its semantic analysis systems can be applied to a billion-page-scale web archive. The software platform has an extendable architecture, and we are planning to enhance WISDOM2013 in the future by adding more semantic analysis systems and inference mechanisms.

1 Introduction

The range of questions is unlimited that humans can pose, and web texts are a valuable information source for finding a comprehensive list of answers, which may include "unknown unknowns" in the infamous words of D. H. Rumsfeld: things that "we don't know we don't know" (Torisawa et al., 2010). However, current commercial search engines are not an effective tool for finding such answers. For instance, even though deforestation is a serious and widely discussed problem, no exhaustive list of answers exists to the question: "What are the consequences if deforestation continues?" We may encounter serious unknown or unexpected consequences in the future. Many documents on the web describe its possible consequences, but only a small portion can be discovered using commercial search engines, because they just provide a huge number of documents that users have to read. Our ultimate goal is to solve such problems by developing deep semantic analysis technologies, which can provide a list of the possible consequences of deforestation, for instance, and a software platform on which semantic analysis technologies can be applied to a billion-page-scale web archive.

We introduce WISDOM2013, our large-scale web information analysis system that consists of deep semantic analysis systems, including a factoid QA and a non-factoid QA, such as a what-happens-if QA, which answers "What happens if deforestation continues?" and sentiment/information sender analysis. We also introduce the underlying architecture of the software platform, which is designed to process/store two billion web documents and works as a common software platform for various semantic analyses.

NICT previously proposed an information analysis system called WISDOM[1](Akamine et al., 2009), which is a predecessor of WISDOM2013. But the source of its analyses were limited to 100 million web pages, and it did not provide QA services. In addition, the depth and the scale of its semantic analysis was quite restricted because it performs the semantic analysis online after receiving user requests. In contrast, most semantic processing that runs on WISDOM2013 is done offline. WISDOM2013 immediately analyzes each web document after it is crawled and can store the basic analysis results for billions of documents owing to its software platform. Therefore, we can drastically improve the breadth and depth of semantic analyses.

2 Script Outline

In this section, we introduce the major features that we will demonstrate. They exploit the common fundamental analysis results, which are produced by the underlying architecture for large-scale analysis as shown in Section 3.

*Current address: Nuance Communications, Inc., Germany. stijn.desaeger (at) nuance.com

[1]http://wisdom-nict.jp/

The Companion Volume of the Proceedings of IJCNLP 2013: System Demonstrations, pages 45–48,
Nagoya, Japan, 14-18 October 2013.

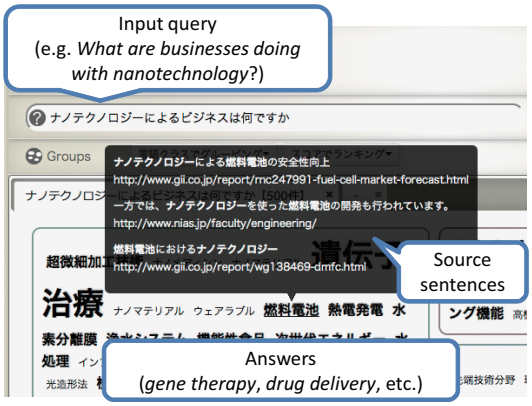

Figure 1: WISDOM2013 interface: factoid QA.

2.1 Factoid QA

Figure 1 shows WISDOM2013's web browser interface. WISDOM2013 takes a question in natural language and returns answers. For example, given question *"What are businesses doing with nanotechnology?"* (ナノテクノロジーによるビジネスは何ですか), WISDOM2013 returns hundreds of answers, such as *gene therapy* (遺伝子治療), *drug delivery* (ドラッグデリバリー), and *artificial joints* (人工関節) and displays them in clusters of semantically related terms. Users can click on each answer to see the original sentence and document from which the answer was extracted.

The system extracts such patterns as "X are businesses doing with Y" in questions and automatically paraphrases the extracted patterns into many synonymous patterns (De Saeger et al., 2009). Those patterns are matched against the web texts using specially designed indexes.

Note that we aim to provide a wide range of answers to user questions, unlike such traditional factoid QAs as IBM's Watson for the Jeopardy! game show (Ferrucci et al., 2010), and suggest unexpected information to users, in other words, "unknown unknowns" (Torisawa et al., 2010). We expect that such unknown unknowns broaden thought and trigger proper decision makings in users.

This technology is an extension of the one used in our voice-activated open domain question answering system (Varga et al., 2011). When it is given a question, "What part of Japan was previously hit by tsunamis?", it found that the Sendai plain, which was devastated by a huge tsunami in the Great East Japan Earthquake in 2011, was also hit 1,000 years ago by a huge tsunami; tsunamis of similar scale are expected to hit again in the future. The system found this answer *from web pages posted before* the Great East Japan Earthquake. For a large number of victims of the 2011 tsunami, this is an example of an "unknown unknown" (or at least relatively unknown facts), and if it had been more widely circulated, lives might have been saved. WISDOM2013 will give chances for many users in the future to discover such *unknown unknowns*.

2.2 What-happens-if QA

When an input question follows the "What happens if X" pattern, WISDOM2013 invokes a special type of QA system, which we call *What-happens-if* QA, and gives the result as a directed graph (Fig. 2). The graph represents the causal chains initiated by the event described in the question. If the given question is *"What happens if deforestation continues?"* (森林破壊が続くとどうなる？), then WISDOM2013 gives a graph that includes the causal chains initiated by the event, "deforestation continues". For instance, the graph contains the following causal chain: "deforestation continues" → "global warming progresses" → "sea temperature rises" → "Vibrio parahaemolyticus swells."[2].

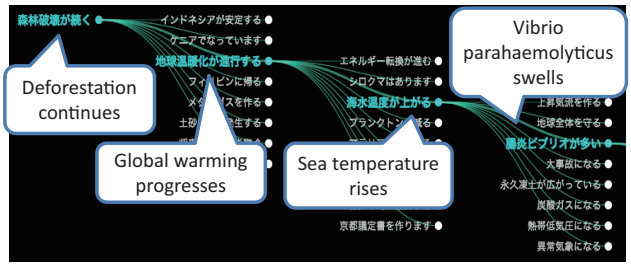

Figure 2: *What-happens-if* QA

Of course, it is debatable whether such causal chains or *future scenarios* will actually happen, and many scenarios are unlikely to become true. Our aim is to provide users the *big picture* of the future concerns of a given question, which is unlikely to be covered by journalism or mass media. We expect that careful examination of such future scenarios will lead to better decision making and preparation for potential and unforeseen risks.

Note that the causalities among nodes are acquired by our previous method (Hashimoto et al.,

[2] An article in *Nature Climate Change* reported that Vibrio infections are caused by global warming in the Baltic Sea (Craig Baker-Austin et al., Nature Climate Change, Vol. 3, pp, 73–77 (2013))

2012) from a large body of web texts. Each single causal relation between two nodes is extracted from a single web page, but a chain of causalities is obtained by combining those extracted causal relations and represents the information scattered over many web pages. In this sense, *what-happens-if* QA involves an certain inference process and enables users to explore possible social scenarios by chaining/combining statements from different documents. In other words, this feature creates awareness of hypothetical future scenarios that are not actually written in any document.

2.3 Sender/Sentiment Analysis

WISDOM2013 can also show the results of sentiment analysis for a given topic or answers for factoid QAs. The amount of positive/negative information based on sentiment analysis is shown in charts to elucidate trends for users (Fig. 3). The results are classified based on the types of senders of the information source page. These functionalities were inherited from WISDOM, WISDOM2013' s predecessor (Akamine et al., 2009). For instance, we can check the reputation of treatments for atopic diseases by applying sentiment analysis to the answers to *"What works for atopy?"* (アトピーに効くのは何ですか) and use the results as clues for determining the treatment' s reliability or uncovering side-effects. In our demonstration, we show that companies post the most positive opinions concerning nutritional supplements that are supposedly effective against atopy. Users might infer that the companies are exaggerating the drug's positive qualities even though much positive information is available about them. In extreme cases, users may question the effectiveness of such supplements or associate side-effects with them.

3 Software Platform

In this section, we describe the architecture of the underlying software platform, which consists of two stages of data processing. Fig. 4 shows the first stage for fundamental analyses and archiving. The fundamental analysis results are designed to be shared by a wide variety of application-oriented analyses in the second stage.

After the crawler collects web documents, fundamental analyses are applied to them, which include document structure analysis, dependency

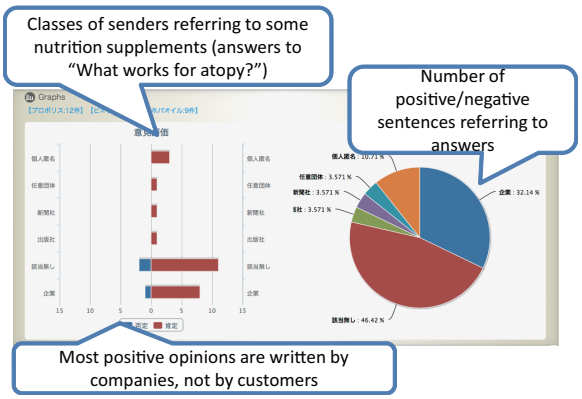

Figure 3: Sender/sentiment analysis

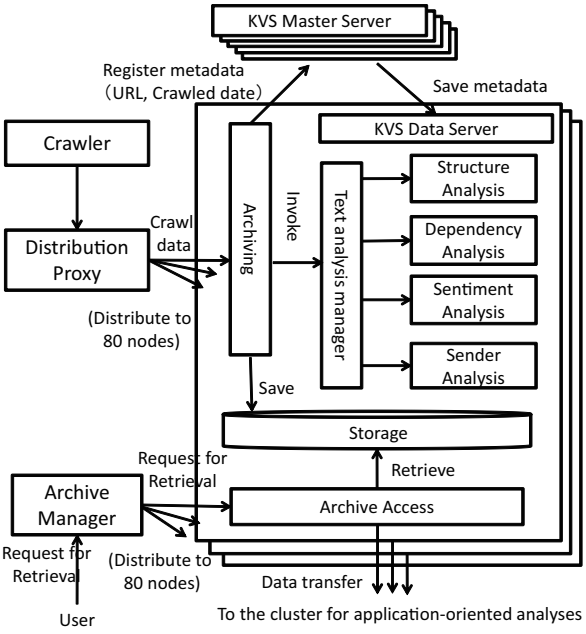

Figure 4: Fundamental analyses and archiving.

analysis, sender analysis, and sentiment analysis. The analyses need to process more than ten million documents daily collected by the crawler. To manage such metadata as URL, the crawled date, and the processing status of each document, we adopted a distributed key-value store (KVS).

In the second stage, more application-oriented analyses are performed based on the fundamental analysis results. For factoid QAs, the preprocessor extracts patterns of phrases that indicate relationships between terms and indexes them. The preprocessor for the what-happens-if QA extracts causal relations and indexes them. Both QAs rely on structure analysis and dependency analysis, both of which are produced in the first stage. Sender/Sentiment are also indexed for the interactive analysis described in Section 2.3. A full text

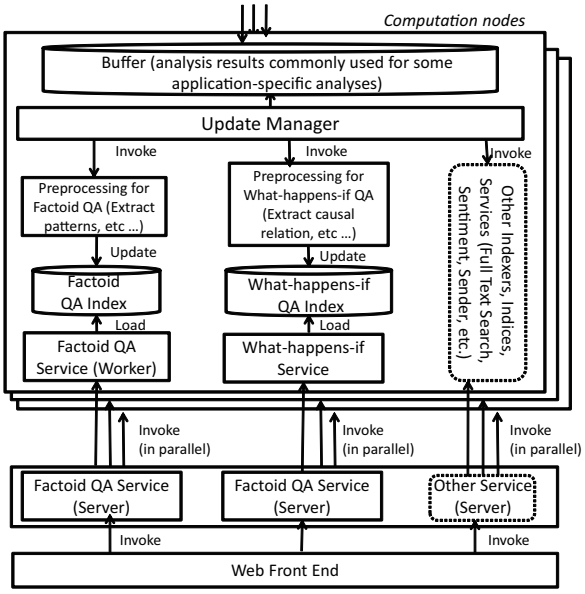

Figure 5: Application-oriented analyses.

search also becomes available based on indexing in this stage. Fig. 5 shows an overview of the process. The update manager transfers the fundamental analysis results to the distributed computation nodes. Due to computational load and data size that exceeds the storage amount of a single node, preprocessing including indexing for some analyses runs on 40 nodes in parallel. The indices for the analyses are used by *access services*, which provide APIs to access the indices. The distributed services are called *worker* services. A *server* service receives a request from the GUI, sends it to all *worker* services in parallel, and aggregates their results. The *server* service also eliminates duplicated results and ranks them. The extensible software platform allows us to add new preprocessors, indices, and services.

4 Conclusion

In this paper, we introduced the major features of WISDOM2013 and described its software platform. We are planning to extend it in the future by adding more semantic analysis systems, such as Why QA (Oh et al., 2013) and inference mechanisms (Tsuchida et al., 2011). We also plan to introduce WISDOM2013 as infrastructure for a counter disaster information analysis system (Ohtake et al., 2013), which we are developing to organize information extracted from tweets after disasters (Varga et al., 2013). WISDOM2013's software and service are scheduled to be made public in 2014.

References

Susumu Akamine, Daisuke Kawahara, Yoshikiyo Kato, Tetsuji Nakagawa, Kentaro Inui, Sadao Kurohashi, and Yutaka Kidawara. 2009. WISDOM: A web information credibility analysis system. In *ACL/AFNLP 2009 (Software Demonstrations)*, pages 1–4.

Stijn De Saeger, Kentaro Torisawa, Jun'ichi Kazama, Kow Kuroda, and Masaki Murata. 2009. Large scale relation acquisition using class dependent patterns. In *ICDM'09*, pages 764–769.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A. Kalyanpur, Adam Lally, J. William Murdock, Eric Nyberg, John Prager, Nico Schlaefer, and Chris Welty. 2010. Building watson: An overview of the DeepQA project. *AI Magazine*, 31(3):59–79.

Chikara Hashimoto, Kentaro Torisawa, Stijn De Saeger, Jong-Hoon Oh, and Jun'ichi Kazama. 2012. Excitatory or inhibitory: A new semantic orientation extracts contradiction and causality from the web. In *EMNLP-CoNLL 2012*, pages 619–630.

Jong-Hoon Oh, Kentaro Torisawa, Chikara Hashimoto, Motoki Sano, Stijn De Saeger, and Kiyonori Ohtake. 2013. Why-question answering using intra- and inter-sentential causal relations. In *ACL 2013*, pages 1733–1743.

Kiyonori Ohtake, Jun Goto, Stijn De Saeger, Kentaro Torisawa, Junta Mizuno, and Kentaro Inui. 2013. Nict disaster information analysis system. In *IJCNLP 2013 (Demonstration Track)*.

Kentaro Torisawa, Stijn de Saeger, Jun'ichi Kazama, Asuka Sumida, Daisuke Noguchi, Yasunari Kakizawa, Masaki Murata, Kow Kuroda, and Ichiro Yamada. 2010. Organizing the web's information explosion to discover unknown unknowns. *New Generation Computing (Special Issue on Information Explosion)*, 28(3):217–236.

Masaaki Tsuchida, Kentaro Torisawa, Stijn De Saeger, Jong Hoon Oh, Jun'ichi Kazama, Chikara Hashimoto, and Hayato Ohwada. 2011. Toward finding semantic relations not written in a single sentence: An inference method using auto-discovered rules. In *IJCNLP 2011*, pages 902–910.

István Varga, Kiyonori Ohtake, Kentaro Torisawa, Stijn De Saeger, Teruhisa Misu, Shigeki Matsuda, and Jun'ichi Kazama. 2011. Similarity based language model construction for voice activated open-domain question answering. In *IJCNLP 2011*, pages 536–544.

István Varga, Motoki Sano, Kentaro Torisawa, Chikara Hashimoto, Kiyonori Ohtake, Takao Kawai, Jong-Hoon Oh, and Stijn De Saeger. 2013. Aid is out there: Looking for help from tweets during a large scale disaster. In *ACL 2013*, pages 1619–1629.