

# Mitsuo Yoshida

Department of Computer Science,  
Graduate School of Systems and Information Engineering,  
University of Tsukuba

—  
Tennodai 1-1-1, Tsukuba, IBARAKI, 305-8573 JAPAN  
+81-29-853-5382 (Laboratory)

Email : ceekz@mibel.cs.tsukuba.ac.jp  
Skype : ceekz\_skype

## Education

Apr. 2011 - Present : **Doctoral Program in Computer Science**

Graduate School of Systems and Information Engineering, University of Tsukuba  
Supervisor : Prof. Mikio Yamamoto

Apr. 2009 - Mar. 2011 : **Master of Engineering, Master's Program in Computer Science**

Graduate School of Systems and Information Engineering, University of Tsukuba  
Supervisor : Prof. Mikio Yamamoto

Master Thesis (in Japanese) : Automatic Extraction for Blog Posts and Comments using a Set of Blog  
**Distinguished Student Award from the Chair**

Apr. 2003 - Mar. 2009 : **Bachelor of Information Sciences, Bachelor's Program**

College of Information Sciences, University of Tsukuba  
Bachelor Thesis (in Japanese) : Primary Content Extraction from Web Pages without Training Data

## Work Experience (including part-time jobs)

Apr. 2011 - Present : **Research Fellow (DC1) - Japan Society for the Promotion of Science**

This fellowship program is granted to doctoral course students who will play an important role in future scientific research activities in Japan.

Mar. 2006 - Present : **Founder - TechTech Inc.**

I started this company in 2006 to put my research into practical use. It provides a search engine (Ceek.jp) and develops customized software. I was the president of this company from Mar. 2006 to Mar. 2011.

Jul. 2011 - Feb. 2012 : **Research Intern - Microsoft Research Asia**

Mentor: Dr. Yuki Arase (Natural Language Computing Group)

Apr. 2009 - Mar. 2011 : **Teaching Assistant - University of Tsukuba**

Taught "Data Structures and Algorithms Laboratory" and "Information Media Laboratory (Developing a Morphological Analyzer)" to undergraduate students.

Sep. 2008 - Mar. 2009, Apr. 2010 - Mar. 2011 : **Visiting Researcher - Tsukuba Industrial Liaison and Cooperative Research Center, University of Tsukuba**

Developed a blog service (hosomichi.jp) which automatically identifies the pathway of the user who made a trip and wrote of it in the blog. The pathway is shown in a map. (Sep. 2008 - Mar. 2009)  
Worked on a research about the next-generation electronic library. (Apr. 2010 - Mar. 2011)

**Apr. 2004 - Sep. 2006 : Software Engineer - Business Search Technologies Corporation**

Developed web crawlers, which gather blog entries and official gazettes automatically.

Developed recommendation modules, which recommend keywords suitable for narrowing down search results.

## Other Projects

**Oct. 2010 - Present : Project Lie ([project-lie.org](http://project-lie.org))**

The purpose of this project is to change “Library and Information Science” into “Library and Information Engineering”. We would like to connect the librarians and LIS researchers and foster “engineers” who are masters of information skills and librarianship. Every Friday, we broadcast the latest news and discussions on libraries by using Ustream.

**Aug. 2002 - Present : Ceek.jp**

This project had been a personal project since 2002, and transferred to TechTech Inc. in 2006.

Provides a meta search engine ([www.ceek.jp](http://www.ceek.jp)). This service unifies the results of multiple search engines and displays it.

Provides a robot type news search engine ([news.ceek.jp](http://news.ceek.jp)). This service is one of the first Japanese news search engine using robots.

## Publications (Selected)

### *Journal Papers*

Mitsuo Yoshida, Asuka Matsumoto. The Use of Social Media in Politics –Applications and Analyses–. *Journal of Japanese Society for Artificial Intelligence* (in Japanese), val.27, no.1, pp.43-50, 2012.

Mitsuo Yoshida, Mikio Yamamoto. Primary Content Extraction from News Pages without Training Data. *DBSJ Journal* (in Japanese), val.8, no.1, pp.29-34, 2009.

### *Conference Papers*

Mitsuo Yoshida, Takashi Inui, Mikio Yamamoto. Automatic Generation of Future Timeline using Web News Corpus. *Processing of the 3rd Rakuten R&D Symposium* (in Japanese), 2010.

Mitsuo Yoshida, Takashi Inui, Mikio Yamamoto. Automatic Generation of Rules on CSS Selector of Primary Content. *Processing of the Rakuten R&D Symposium 2009* (in Japanese), pp.7-10, 2009.

## Awards

**May. 2011 : Exemption from Repayment by Particularly Outstanding Results**

Mitsuo Yoshida. Japan Student Services Organization.

**Mar. 2011 : Distinguished Student Award from the Chair**

Mitsuo Yoshida. Department of Computer Science, Graduate School of Systems and Information Engineering, University of Tsukuba.

**Dec. 2010 : Best Data Challenger Award**

Takaaki Tsunoda, Kento Sawada, Mitsuo Yoshida. 3rd Rakuten R&D Symposium.

Nov. 2009 : **Excellent Paper Award**

Mitsuo Yoshida, Takashi Inui, Mikio Yamamoto. Rakuten R&D Symposium 2009.

Mar. 2009 : **Excellent Interactive Award**

Mitsuo Yoshida, Mikio Yamamoto. The First Forum on Data Engineering and Information Management.

## Computer Skills

Languages : Perl, PHP, C#, C, (D)HTML, JavaScript, CSS, SQL  
9+ years of Web application development experience, including LAMP architecture.

## Research Experience

Sep. 2010 - Present : **Automatic Generation of Future Timeline using Web Pages**

Future prediction is necessary for people and companies to develop strategies for the future. In addition, people basically enjoy making predictions for fun.

We proposed methods to extract sentences including future events from a collection of Web news and generate "Future Timeline" automatically using the sentences to get a quick overview of the future.

Apr. 2008 - Present : **Primary Content Extraction from Web Pages**

In recent years, the proportion of primary content in a Web page has been decreasing as content management systems (CMS's) continue to spread, because CMS's automatically and excessively add unnecessary parts such as menus, advertisements and copyright notices into the Web page.

We proposed a simple method extracting the primary content from a collection of Web pages without training data. We regard a Web page as a set of blocks (minimum unit of primary or non-primary content), and assume that blocks of the primary content are unique and there are copies of those of non-primary content. Additionally, we proposed a simple method to generate special CSS selectors as rules of extracting primary content from a collection of Web pages.

We showed that the proposed method can accurately extract the primary content from real pages of blog and news sites.

Sep. 2009 - Mar. 2010 : **Analysis of Tweets including URLs on Twitter**

Web 2.0 strengthened both tropism of the information to be able to put on Web and brought about new Web service. The microblogging service including Twitter is nominated for a kind of the Web 2.0 Web service. The microblogging service functions as a sort of personal information platform of the online social network service.

We paid our attention to tweets including URLs on Twitter, which is representative of the microblogging service, and investigated the characteristic.

As a result, we showed that news sites with much number of the unique users does not attract attention on Twitter, and the number of retweets is not even remotely related to the number of URLs posted, and so on.