Information Credibility Analysis of Web Content

Yutaka Kidawara

Group Leader
Knowledge Clustered Group.
Knowledge Creating Communication Research Center
Keihanna Laboratory
NICT

National Institute of Information and Communications Technology
What is NICT?

NICT is only one governmental laboratory for communications and information technology in Japan.

Message from the President:

In April 2004, the Communications Research Laboratory, an incorporated administrative agency, and the Telecommunications Advancement Organization of Japan, a chartered corporation, were merged and re-launched as the National Institute of Information and Communications Technology (NICT), an incorporated administrative agency.

NICT was established to carry out research and development in the field of information and communications technology, in an integrated manner from basic science to application, with the aim of supporting the approaching ubiquitous network society as well as to provide comprehensive assistance to public and private organizations working in this field.

In order to fully realize the merits of the merger, our first undertaking was to create a new system of cross-departmental research and development promotion units by organically combining the functions of our two predecessors. NICT is now coming into the final year of the medium-term plan that was launched by CRL in 2001. We are working day and night to provide the Japanese public with visible evidence of the results of all the research and development work that has taken place during the past five years.

We are pursuing our research and development activities from a long-term perspective with our eyes set firmly on a paradigm for the future of information and communications, in other words, our vision for the future. This strategy links into our next medium-term plan, which will start next year.

Information and communications technology is the field of technology that provides a platform for all industrial activity. NICT will continue to provide technological support in the formulation of government policies.
About Us (Knowledge Clustered Group)

Mission

- Research and development of the knowledge processing technology to support human's intelligent activities in network society.
- Development of information analysis technology for universal communication.

Research projects (April 2006 – March 2011)

- Knowledge Cluster project
  - Information Credibility Criteria project
    - Budget: Approx. 2.5 million $/year

NICt and MIC(*) contract projects (April 2007 – March 2011 )

- (A) Information analysis technology for Web content
- (B) Information analysis technology for semantics on Web content
- Budget: Approx. 2.5 million $/year

*Ministry of Internal Affairs and Communications
Searching Web pages by keywords.

I want to take care of my health. What is Healthy food “Agaricus”? Is it effective?

The page describes “Agaricus involves various minerals. It’s very good for your health”

“A professor finds healthy people who are less likely to fall ill. They eat often Agaricus”

It seems to be good!!
We cannot find credible information due to huge amount of search results.

Search result of “Agaricus”

Advertisement saying it’s good for your health! But, is it really true?

We need more multiple information perspectives to find credibility!

- What’s the affiliation of page owners?
- Are contact address and info sources provided?
- How good is their reputation?
- Are there contradictory statements?

Information analysis technology different from conventional search engines is essential!
How can we find credible information?

- We need various information perspectives to find credible information on Web pages.

A variety of methods is necessary to analyze various content types.
NICT and MIC Information Credibility Analysis Project

- NICT is developing Information Analysis engine for Web Content as an own research project (WISDOM).
- MIC (Ministry of Internal Affairs and Communications) and NICT fund two research projects (Contract research A/B).
- All the projects are proceeding collaboratively to tackle difficult development.

To Analyze more variety of digital contents
- Analysis of
  - elemental digital content on Web pages.
  - Credibility of ordinal search engines
  - Knowledgability of blogger

Contract research (A)
Information analysis technology for Web content (i-Believe Project)
- Kyoto Univ.
- Hyogo Univ.
- Kyoto Sangyo Univ.
- Kizashi

WISDOM
NICT’s own project

Information Analysis Technology for Information Credibility on the Web

Create new research area of the computer science!!

To analyze more deeply text content
- Analysis of temporal changing opinions
- Identify the logical relations relevant to the user’s query

Contract research (B)
Information analysis technology for semantics on Web content
- NEC
- NAIST: Nara Advanced Institute of Science and Technology
- Yokohama National University

TSUBAKI
Kyoto Univ.
MEXT info-plosion project
NICT Information Credibility Criteria Project

Web Information Analysis System (WISDOM)
How can we find credible information?

We need various information perspectives to find credible information on Web pages.

- What is NICT?
- Is the picture appropriate as this page?
- When did he become the president?
- The building is there still now?
- Is the opinion positive or negative?
- What kinds of opinions are on the Web?
- How much knowledgable is writer/blogger?
- Does the page have proper appearance?
NICT Information Credibility Criteria Project

Development of an information credibility analysis system which supports human judgments by collecting, analyzing, and organizing related information automatically.

1. Computers collect information
   - The Internet
   - a mixture of good and bad content on the Web

2. Computers organize information
   - Information analysis by WISDOM
   - Check the contact address and site policy
   - Identify the information senders and their role
   - Extract the evaluative expression (reputation, opinion,...)
   - Grasp the major/contradictory statement

3. Computers help human being judge information credibility
   - Sender
   - Reputation: Positive
   - Appearance: Good
   - Sender: NICT

NICT Information Technology
What is the WISDOM?

- WISDOM (Web Information Sensibly and Discreetly Ordered and Marshaled) analyzes Web content.
Functions of WISDOM

Information credibility criteria

Appearance
- Check the contact address and site policy

Information analysis by WISDOM

Sender
- Identify the information senders and their role

Content
- Extract the evaulative expression (reputation, opinion,...)
- Grasp the major/contradictory statement
Appearance Information

Appearance information = document style and superficial characteristic

Information extraction
An information extraction module extracts information which related to the credibility of Web page.

- Whether or not the contact address (organization’s physical address, contact phone number, contact email address) is shown in the page
- Whether or not the privacy policy is in the page.
- The volume of advertisements
- The sentential style (spoken-language or not)

Web page classification
A Web page classification module filters out some pages for information analysis, and also assigns a basic page-type

- Unselected pages:
  - Web pages that do not involve enough sentences to analyze (e.g. Search result page, Links page, Sales page, Photo album page)
  - Selected page (Basic page-type)
    - Blog, Bulletin board, ordinal Web pages

Which is more credible?

Privacy policy is shown in the page
Contact address
No advertisement

No organization’s physical address
No Privacy/Site policy
Many advertisements (affiliate links)
Appearance Information on WISDOM analyzed lists

Details of Analysis

- Number of sentences (noun phrase, sentence)
- Number of images (large, medium, small)
- Number of links (inside, outside)
- Number of advertisement
- Date

Blog | Contact Address | Volume of Ad
Functions of WISDOM

Information credibility criteria

Appearance
Check the contact address and site policy

Information analysis by WISDOM

Sender
Identify the information senders and their role

Content
- Extract the evaluative expression (reputation, opinion,...)
- Grasp the major/contradictory statement
Extraction of Evaluative Information (1/2)

- Evaluative expression
  - I like apples.  
    (Explicit subjective expression)
  - This camera was broken 3 days after I bought it.  
    (Implicit objective expression)

We handle both subjective and objective evaluative expressions

- Evaluative expression types
  - Evaluation (+/-) : Approval/disapproval or praise/criticism
  - Emotion (+/-) : Human feeling
  - Merit (+/-) : Merit/demerit
  - Event (+/-) : God/bad events or situations
  - Adoption (+/-) : Adoption or promotion
  - Deontic : Proposal, advice, hope or request
A corpus tagged with evaluative information was constructed for developing the evaluative information extraction system.

6,000 sentences were collected from the Web for 30 topics such as bio-ethanol and pension plan.

Evaluative expressions in sentences, their holders, types and sentiment polarities were annotated.
Functions of WISDOM

Information credibility criteria

Appearance
Check the contact address and site policy

Information analysis by WISDOM

Sender
Identify the information senders and their role

Content
- Extract the evaluative expression (reputation, opinion,...)
- Grasp the major/contradictory statement
Information Sender Analysis

- The source of information (information sender) is one of the important elements for judging the credibility of information.
- Questions related to the information sender that the user might ask:
  - Who wrote this page?
  - What credentials or expertise does the author have?
  - Can I trust the Web site/author?

“Extracting the Author of Web Pages"
Yoshikiyo Kato (NICT, Japan), Daisuke Kawahara (NICT, Japan), Kentaro Inui (NAIST, Japan), Sadao Kurohashi (NICT and Kyoto University, Japan), Tomohide Shibata (Kyoto University, Japan)

WICOW2008 session 2
(after my talk)
Research on Web Content Analysis Technologies
NICT and MIC contract research project (A) : Information analysis technology for Web content (i-Believe Project)

Kyoto University, University of Hyogo
Kyoto Sangyo University, Kizashi Company
(In cooperation with Yahoo! Japan)
How can we find credible information?

- We need various information perspectives to find credible information on Web pages.

- What is NICT?
- Is the opinion positive or negative?
- What kinds of opinions are on the Web?
- How much knowledgeable is writer/blogger?
- Does the page have proper appearance?
- Is the building there still now?
- When did he become the president?
- Is the picture typical one?
Objective

• Explore technologies to help people evaluate the information credibility of Web contents (images, sounds, video & text) returned by Web search engines etc.
• Non-deep analysis of the target contents and their related Web contents
• ‘Reasonable’ processing time
• Experimental evaluation through real ISP environment

Goal

• Precision ratio: more than 80%
• Processing time:
  Less than 10 seconds (Real-time analysis),
  Less than 10 minutes (On-demand type analysis)

Approach

• Analysis of Content itself
• Analysis of Support for Content
• Analysis of Information Disseminators
Results & Current Status (1)

Web Image Credibility (Kyoto University)

- Analyze how a target image is different from real world images or other related images (typicality and speciality analysis)
- Analyze the ‘coverage’ ratio of a target’s image set to evaluate the target ‘object’ by searching and gathering complementary images from other resources and by computing the coverage rate of the target image set
- Analyze the consistency of people’s relationship who co-appear in a Wikipedia image by gathering related people information from Wikipedia & by computing the strength of the people’s relationship by Wikipedia hyperlink structure

Web Video Credibility (Kyoto University, University of Hyogo)

- Analyze video’s bias by gathering similar & complementary contents from other resources and by showing the distribution
- Analyze the video’s credibility by analyzing viewers’ annotation by positive/negative & sentiment analysis technologies
- Extract disseminator’s intent by analyzing both of his/her disseminated & non-disseminated contents

Web Digital Map Credibility (University of Hyogo)

- Validation of map credibility using co-occurrence relations between geographical objects on the Web archive
Web Text Credibility (Kyoto University)

• Explored light-weight analysis of Web page using search engine indices. The developed system evaluates the fairness of the searched pages (whether it is a majority and how much it contains typical topics concerned with a query term) and the degree of social support for the searched pages (how many bookmarks are attached to the page, and how widely and uniformly the page is supported (hyperlinked)) geographically.

Web Contents Disseminator Credibility (Kyoto Sangyo University)

• Explored analysis techniques to evaluate the sentiment of content disseminator (sentiment analysis) and how much the content disseminator knows well about the domain he/she discusses (knowledgeability analysis).
• Developed a ‘sentiment dictionary’ by analyzing news article 6-year archive (Yomiuri and Asahi newspapers), which gives sentiment degree for each term that frequently co-occurs together with typical expression terms.
• Developed a way to compute the knowledgeability of bloggers by analyzing his/her blog data.

Blogger Classification and Blog Ranking by Bloggers’ Knowledgeability (Kizashi Company)

• Together with Kyoto Sangyo University, we explored a ranking method based on blogger’s knowledgeability by creating a Knowledge Domain Dictionary.
• Also, we started to evaluate the developed method in a Kizashi Lab site.
Multimedia Credibility
(Web movies, sound & images)
Video Movies on Web pages

How much credible is the cooking video on web page?

Is the video news biased compared with other info sources?

Analysis of the video's credibility by viewers' annotation
- positive/negative & sentiment analysis

Evaluate the differences between the closed captions of the target video & other source video
- Search for and analyze related video & audio & Web contents from other source.
- Show the distribution of topics and opinions
Credibility of Multimedia: Cross-media Analysis of Video Consistency

Detecting Real-world Clipping of Video

Search for related information (video clips and web pages) of a given video clip from other media or source by using its closed caption and surrounding text, and discover the viewpoint difference in video descriptions* by the comparison of related information. Clarify the bias and diversity in video contents to support the credibility evaluation of users.

① Search for related information based on complementary information retrieval mechanism

② Cluster the related information to analyze and present the description difference in different media

N station reported that France would boycott the open ceremony of Beijing Olympic Games. However, other stations report that EU would boycott open ceremony of Beijing Olympic Games rather than France.
Multimedia Credibility: Analysis Video Clips Based on Social Annotation (Kyoto Univ.)

System enables users to judge trustworthiness of video clips based on social annotation

- We focused on social annotations of a video clip
- System analyzes comments which are posted by viewers and divides them into positive, negative and some sentiments by using dictionary
- System displays the changing the level of positive, negative and sentiments by time-related graphs

"Can Social Annotation Support Users in Evaluating the Trustworthiness of Video Clips?"
Satoshi Nakamura (Kyoto University, Japan), Makoto Shimizu (Kyoto University, Japan), Katsumi Tanaka (Kyoto University, Japan)

WICOW2008 session 3 (after lunch)
Pictures on Web pages

Wikipedia
Why does she appear often in a Japanese politician “Nobutaka Machimura” Wikipedia page. Is the photo appropriate?
Analyze consistency of the Wikipedia image by discovering the strength of the two persons’ relationship by Wikipedia hyperlink

Travel guide
The hotel is great!! Is the picture set enough to describe the hotel?
Analyze the target image set’s coverage rate to evaluate the target object
• Search and gather complementary images from other resources
• Evaluate the coverage rate of the target image set

Enterprise
The food looks like very delicious!! Can we buy it really?
Evaluate how different the image is from real object photos (typicality & speciality analysis) gathered from Web
Propose a method using generalized max-flow on “doubled network”, constructed from link structure of Wikipedia by doubling every link.

Nobutaka Machimura has a strong relationship to Condoleezza Rice through Abe, Aso, and Fukuda, so such a picture is in this article.
Multimedia Credibility: 
Comparative Search Engine and Interface for Evaluating Trustworthiness of an Image on a Web page (Kyoto Univ.)

System evaluates typicality and specialty of target image comparing with other sources’ images

- System calculates the score of typicality of an image on an official restaurant Web page comparing with related images in unofficial Web pages
  Difference between target image and real

- System calculates the score of specialty of a restaurants’ menu image comparing with related images in another official Web pages
  Difference between target image in an official site and images in other official sites
Multimedia Credibility: Comparative Search Engine and Interface for Enabling Users to Judge Trustworthiness of Images (Kyoto Univ.)

Search related images from other sources and show them

- System calculates the coverage about aspects of hotels’ images
- Aspects: room, bed, bath and shower, toilet, view, foods, appearance

- System searches complementally images from other sources to increase the coverage of aspects

- System also shows the neighboring images by using location information
Multimedia Credibility: Spatio-Temporal Consistency Analysis for the Map Credibility

(University of Hyogo)
Digital map credibility is very important problem. Digital maps were widely used on the Web. However, digital maps are not frequently updated. For instance, old objects are displayed on digital maps, new objects are not displayed on digital maps, and so on.

In this work, we developed an analyzing method of geographical object existence using metadata of digital maps and web information archives for validation of map credibility.
Web Text Credibility

- Analysis of Content itself
- Analysis of Support for Content
- Analysis of Information Disseminators
Web Text Credibility: Analysis of Content itself and Analysis of Support for Content

Additional information to evaluate text credibility is obtained in real time

Fairness of the Content

Features of the content itself.

- **Topic Coverage of the Content**
  - How much the content contains typical topics.

- **Topic Majority of the Content**
  - How many other contents are similar to the content.

- **Speciality of the Content**
  - How much the content contains specialized information.

Social Support for the Content

Evaluation from the outside of the content.

- **Social Bookmark for the Content**
  - How many times the content is bookmarked.

- **Local Support for the Content**
  - How widely people support (link to) the content (locations).
Web Text Credibility: Light-Weight Analysis of Text using Search Engine Indices (Kyoto-U)

- Showing additional information with Web search results for users to evaluate credibility.

- Fairness of the searched pages
  - Whether it is a majority.
  - How much it contains typical topics.

- Social support of the searched pages
  - How many bookmarks are made.
  - How widely are supporters located.

- Access to the search engine indices

- Extract information to evaluate credibility

Various Search Engines

Indices

Hyperlinks

Social Bookmarks
Web Text Credibility: Light-weight analysis of text using search engine indices (a system developed in Kyoto-U)

Snippet analysis of the snippets

- Topic majority
- Topic coverage
- Social bookmark support
- Local support
- Style of writing

Extension to question-answer services

- QA services
- Forums, BBSs

Ex) Informational query “Kyoto Leisure”:
Topic majority and topic coverage can work well to think about credibility.
Sentiment Analysis and Knowledgeability Analysis of Information Disseminators for Web Information Credibility

Kyoto Sangyo University
Web Information Credibility based on Sentiment analysis

Detecting sentiment about a topic for social background

- Sentiment analysis for each domain such as web news site and country.
- We can provide baseline of each credibility based on sentiment

"Using a Sentiment Map for Visualizing Credibility of News Sites on the Web"

Yukiko Kawai (Kyoto Sangyo University, Japan), Yusuke Fujita (Kyoto Sangyo University, Japan), Tadao Kuma (Chiba Institute of Technology, Japan), Jianwei Zhang (Kyoto Sangyo University, Japan), Katsumi Tanaka (Kyoto University, Japan)

WICOW2008 session 3
(after lunch)
Web Information Credibility: Blogger Knowledgeability Analysis
(Kyoto Sangyo Univ. & kizasi Company, Inc.)
Blogger Knowledgeability Analysis

Bloggers are influential disseminators of information on the web. However, their knowledgeability and credibility vary vastly.

**Ranking Method Based on Blogger Knowledgeability**

1) **Creating a Knowledge Domain Dictionary**
   - Perform a web search for keywords such as “Fan”, “Freak”, “Mania”.
   - Register terms that frequently occur right before the keywords as new domains. “Baseball Fan”, “Anime Mania”, etc.
   - Create a knowledge domain dictionary.

2) **Determining Blogger Knowledgeability**
   - For each blogger, assign a score to each entry based upon the occurrence of words in our previously created dictionary.
   - Aggregate these entry scores to calculate a final blogger knowledgeability score.
Research on Survey Report Generation for Assisting Information Credibility Evaluation
NICT and MIC contract research project (B) : Information analysis technology for semantics on Web content

NEC, NAIST, YNU
How can we find credible information?

- We need various information perspectives to find credible information on Web pages.

  What is NICT?
  - Is the opinion positive or negative?
  - Is the building there still now?
  - When did he become the president?
  - What kinds of opinions are on the Web?
  - How much knowledgable is writer/blogger?
  - Does the page have proper appearance?
Goal: Survey report generation

- Assist the user’s judgment of a source’s credibility by
  1. Finding the logical relations relevant to the user’s query,
  2. Detecting the “stale” information, and
  3. Revealing the overall structure of the arguments to the user

1. Logical relation analysis
2. Time-series analysis
3. Summary generation

Supporting / opposing opinions

Information “staleness”

Summary text / graph

Diesel cars are bad for the environment
Diesel cars are low-pollution

(occurrence)

2005 2006 2007 (date)

Diesel automobiles emit a large amount of Nitrogen Oxide when run at high speeds. In downtown Tokyo, diesel automobiles are to blame for the current Nitrogen Oxide situation.

However, diesel automobiles are gaining a reputation as being good for the environment. For example, diesel automobiles are relatively green and their use deserves consideration for the environment.

Elaboration

Diesel automobiles emit relatively less CO2 emissions, and with improvements in technological innovations, they now meet gas emission standards.

Opposition

Diesel automobiles emit a large amount of Nitrogen Oxide when run at high speeds. Diesel automobiles are relatively green and their use deserves consideration for the environment.

By nature, diesel engines have lower CO2 emissions, and with improvements in technological innovations, they now meet gas emission standards.
Background: Web search engines

Needs

- Essential tools for info. seeking

Problems

- Outputs sequentially ordered
  - Difficult to overview retrieved info.
  - Difficult to find minor opinions

- Snippets generated by word proximity
  - Difficult to judge content credibility

- Time-insensitive
  - Difficult to distinguish stale info.

Are diesel engines good for the environment?
Diesel automobiles are to blame for the current Nitrogen Oxide situation.

Diesel automobiles emit a large amount of Nitrogen Oxide when run at high speeds.

In downtown Tokyo, diesel automobiles are to blame for the current Nitrogen Oxide situation.

Diesel automobiles are relatively green and their use shows consideration for the environment.

By nature, diesel engines have low CO2 emissions, and with improvements in technological they now meet gas emission standards.
Detect the “stale” info. by analyzing trends in the occurrence of relevant statements: the decreasing trend and the increasing trend of the opposite statement show the “staleness”.

Time-series analysis

NOx emission

Diesel cars are bad for the environment

Diesel cars are low-pollution

Ground

2005 2006 2007 (date)

Opposition

- less CO2 emission
- Diesel cars are very popular in Europe

bad for the environment

low-pollution

consider the staleness

bad for the environment

low-pollution
The system generates a summary by presenting the logical relations among statements.

- **Goal**: Development of a way of information presentation that contributes to judging information credibility.

  In order for users to easily grasp statements and the logical relations among them, the system generates a summary by presenting the logical relations among statements.

- **Theme**: Summarization of text information on the Web.

- **Logical relations**: Diesel automobiles emit a large amount of Nitrogen Oxide when run at high speeds.
  - In downtown Tokyo, diesel automobiles are to blame for the current Nitrogen Oxide situation.

  - **However**, diesel automobiles are gaining a reputation as being good for the environment. For example,
    - Diesel automobiles are relatively green and their use shows consideration for the environment.
    - By nature, diesel engines have low CO2 emissions, and with improvements in technological they now meet gas emission standards.
Conclusion

- NICT and MIC are continuing to research “Information Analysis for Credibility on Web pages”.
  - NICT is developing Information Analysis Engine “WISDOM”
  - Kyoto Univ. group is developing various methods to evaluate credibility on a variety of digital contents and Web services.
  - NEC group is developing deep analysis methods for text on Web pages.
- We believe those researches can create a new research area for Web technology.
Thank you!!

kidawara@nict.go.jp