

Extracting the Author of Web Pages

Yoshikiyo Kato, Daisuke Kawahara, Kentaro Inui (NICT)

Sadao Kurohashi (NICT / Kyoto Univ.)

Tomohide Shibata (Kyoto Univ.)

NiCT ICC Project

Information Credibility and the Source of Information

- The source of information (*information sender*) is one of the important elements for judging the credibility of information
 - Anonymous vs. a person with a certain background
 - Business selling health food products vs. medical expert
- Elements of the credibility of information sender
 - Competence: expertise, in a position to know
 - Intention: motivation for disseminating credible information or misinformation
- Before we can analyze these characteristics, we need to identify the information sender
 - Focus of this study

Who is the information sender of this page?

The screenshot shows the NICT website homepage. The NICT logo is circled in red. A navigation menu is at the top. The main content area features a header with the text "NICT (National Institute of Information and Communications Technology)" and a large question: "Which one is the information sender of this page?". Below this is a portrait of the President, Shigeo Miyahara, also circled in red. A blue arrow points from the question to the text "Both!". At the bottom, a red-bordered box contains the text: "However, their roles (and the responsibilities concomitant with them) are not the same." The footer includes the text "2006年4月より新しい5か年の中期計画期間がスタートしました。"

NICT (National Institute of Information and Communications Technology)

Which one is *the* information sender of this page?



理事長 宮原 秀夫

Both!

However, their roles (and the responsibilities concomitant with them) are not the same.

Identifying Information Sender Configuration

- **Goal:**
 - To identify the information senders and their roles in publishing the information as information sender configuration
- **Elements of Information Sender Configuration**
 - Information Sender (e.g. “NICT”)
 - Sender Class (e.g. Government)
 - Configuration Type (e.g. bunch)
- **Example**
(bunch ,
 (Government , “NICT”) ,
 (- , “Miyahara Hideo” , “President” , -))

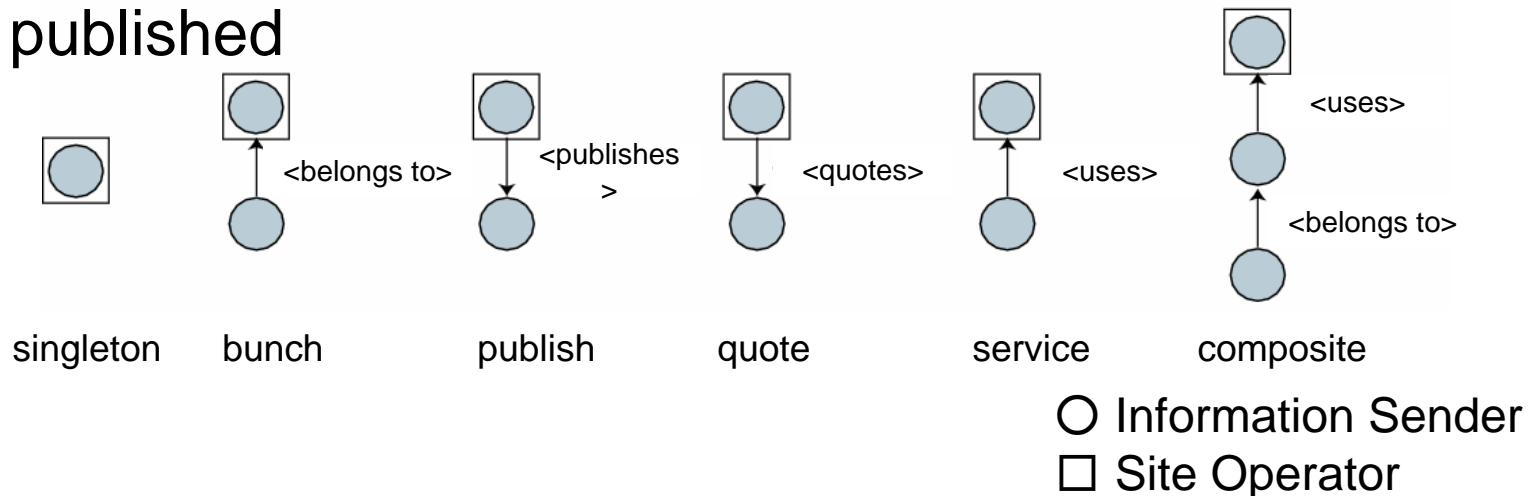
Sender Class

- Information senders are categorized into *sender classes*
- Axes of classification:
 - Individual vs. Organization
 - Profit vs. Nonprofit Orgs.
 - Expertise (Univ., Medical, etc.)
 - Social function (Press)
 - Anonymous vs. Real Name

1. Organization
 - (a) Profit Organization
 - i. Company
 - ii. Industry Group
 - (b) Nonprofit Organization
 - i. Academic Society
 - ii. Government
 - iii. Political Organization
 - iv. Public Service Corporation, Nonprofit Organization
 - v. University
 - vi. Voluntary Association
 - vii. Education Institution
 - viii. Medical Institution
 - (c) Press
 - i. Broadcasting Station
 - ii. Newspaper
 - iii. Publisher
2. Individual
 - (a) Real Name
 - (b) Anonymous, Screen Name

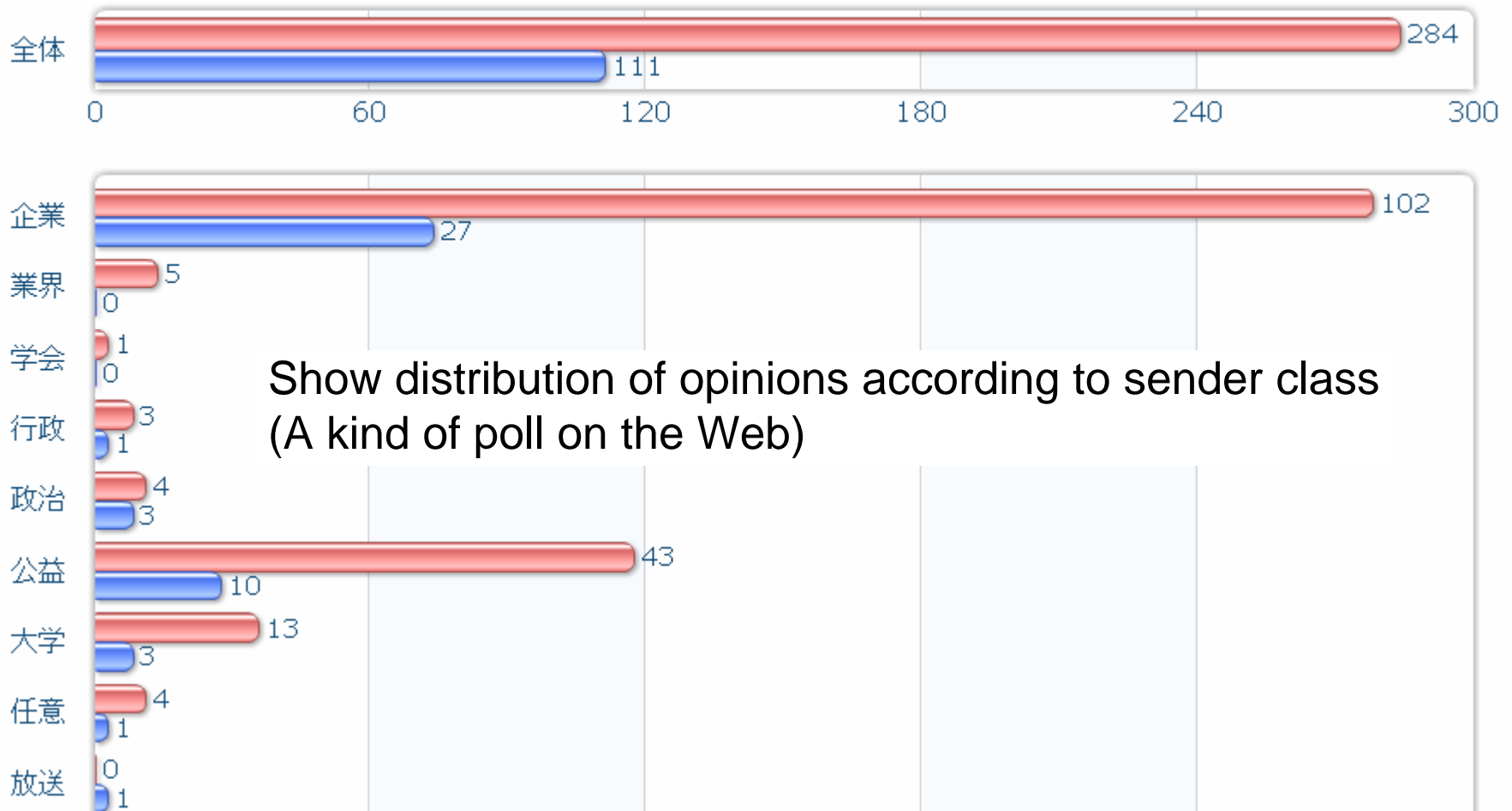
Configuration Type

- Configuration type indicates the relationship between information senders
- Six types have been defined
 - The difference among types are concerned with the attribution of responsibility for the information published



How ISCs are Used?

発信者クラスごとの肯定・否定意見

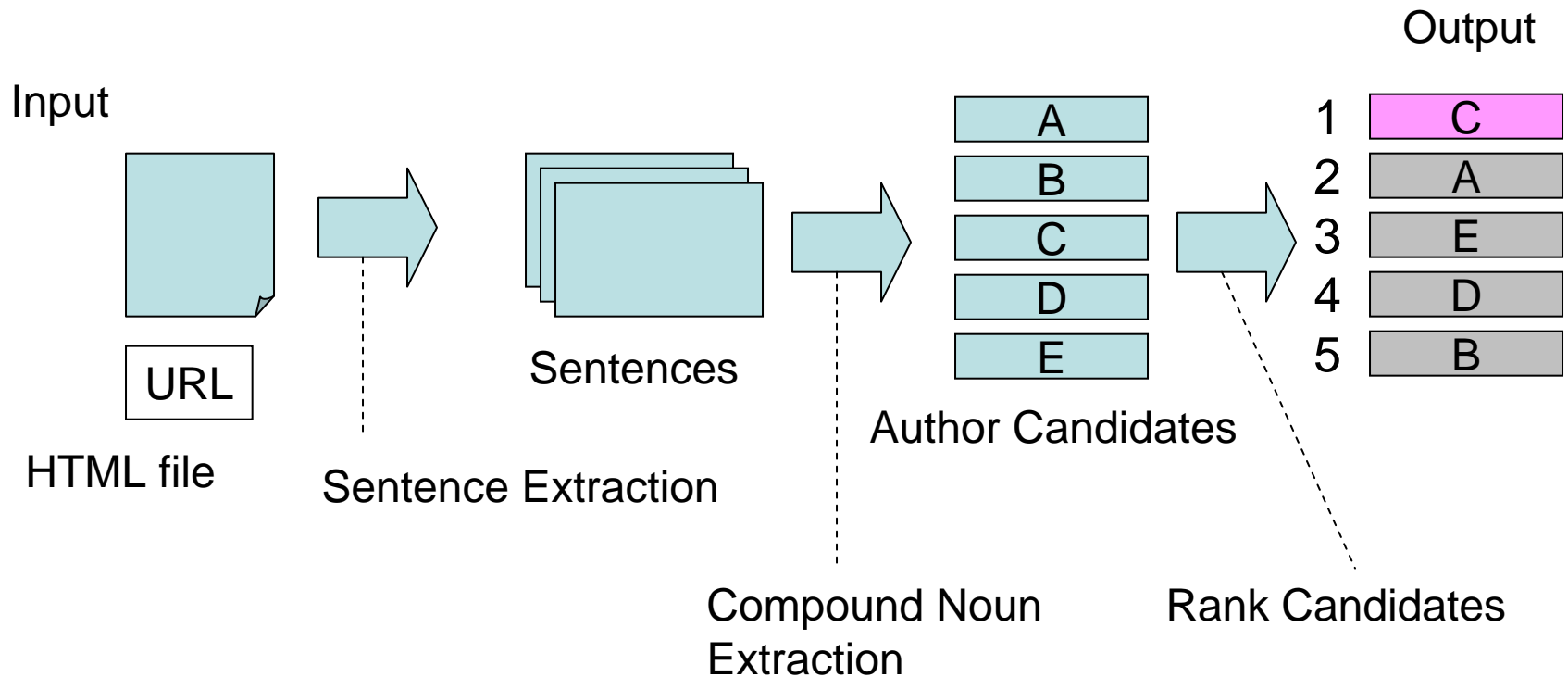


Show distribution of opinions according to sender class
(A kind of poll on the Web)

Toward Identifying Information Sender Configuration (ISC)

- Undertakings so far:
tackling sub-problems of ISC identification
 1. Site operator identification
(e.g. identifying “NICT” as *the site operator*)
Precision: 70%
(N.B.: achieves 80% when blogs/forums excluded)
 2. Sender classification
(e.g. classifying “NICT” as Government)
Precision: 74%
 3. Author identification
(i.e. identifying “Hideo Miyahara” as *the author*)
Precision: 59%

Process of Author Extraction



Sentence Extraction

- Extract text from HTML
- Segment texts into sentences
- Apply Japanese parser (KNP) to the sentences
- Retain sentences that satisfy the following conditions (to filter out *normal* sentences):
 - The ratio of the number particles other than ‘*no*’ (“of”) to the number of morphemes in the sentence is less than a certain threshold
 - Does not contain compound particles with verb (e.g. ‘*~ ni tsuite*’ (“concerning ~”))

Compound Noun Extraction

- Extract clauses that satisfy the following conditions:
 - At least one named entity typed either as a *person* or an *organization* is recognized (using NER functionality of KNP)
 - The sentence contains one or more morphemes typed as *not-in-dictionary word*
 - The last morpheme but particles has a POS tag of *person name suffix* or *organization name suffix*
- Extract compound noun from qualifying clauses

Example of author extraction

URL: <http://www.nict.go.jp/about/message.html>

Title: NICT 独立行政法人 情報通信研究機構

NICT 独立行政法人 情報通信研究機構
NICT 独立行政法人 情報通信研究機構
法人情報
News & Reports
NICTチャンネル
about NICT
NICT憲章
NICTの紹介
NICTのビジョン
NICT沿革
理事長 宮原 秀夫
Copyright National Institute of Information and Communications Technology.

理事長 あいさつ

情報通信研究機構(NICT:National Institute of Information and Communications Technology)は、来るべきユビキタスネット社会を支える情報通信技術の研究開発を、基礎から応用まで一貫した統合的な視点で行い、併せて情報通信分野の事業支援等を総合的に行う独立行政法人です。

私たちは、平成18年4月より、新しい5か年の中期計画期間をスタートしました。この大きな節目を迎えるに当たり、これまで取り進めてきた研究開発事業を「新世代ネットワーク構築技術」「ユニバーサルコミュニケーション基盤」とともに、これらの研究開発を推進す

Author Candidates

情報通信技術(ICT)は、全ての産業活動を支える基盤となる技術分野です。私たちは、ユニバーサルコミュニケーションという概念を掲げ、世界中の人達相互間で自由にコミュニケーションができるようにするとともに、人と機械、機械相互間においても自由にやりとりができる理想の社会の実現を目指して努力しております。

私たちの研究開発成果は国際的な標準化や産業界への技術移転についでゆきます。また、技術の実用化に向けた産学結集型研究開発や、大学・企業等への研究委託、技術の事業化を加速するためのベンチャー支援やインフラの高度化支援など、幅広い活動も行ってまいります。

このように、NICTは、情報通信分野における国の唯一の研究機関として、国の情報通信政策を技術的側面から支えるとともに、大学や産業界、さらには海外の研究機関と密接に連携し、また研究成果の社会への普及に積極的に取り組むことにより、活力ある社会、豊かな生活の実現に向けて努力してまいります。

各所在地一覧
組織
2006年4月より新しい5か年の中期計画期間がスタートしました。

Ranking Author Candidates

- Rank author candidates according to how likely they are indeed the author of the page
- Model: Ranking SVM (Joachims 2002)
- Labels: scores based on the match of the candidate to the human annotation
 - Exact match: 2
 - Partial match: 1
 - No match: 0
- Features
 - TF
 - POS tags (especially, # of PERSON, LOCATION, ORGANIZATION words)
 - Tokens (separate features for the starting and ending tokens)
 - HTML tags
 - Context words
 - Distance from the main content

Positions of the Author Name and the Main Content


The screenshot shows the NICT website's 'about' page. The header includes the NICT logo and navigation menus. The main content area features a portrait of the Chairman, Aizawa, and a text block. A red circle highlights the author's name '理事長 宮原 秀夫' (Chairman Hideo Miyahara) in the text. A line connects this name to a box labeled 'Author Name'. Another line connects the main text block to a box labeled 'Main Content'. A large text box at the bottom contains the hypothesis: 'Hypothesis: Author names appear in the proximity of the main content of the page'.

NICT 独立行政法人 情報通信研究機構

NICTについて ▼ 部門紹介 ▼ 法人情報 ▼ News & Reports ▼ NICTチャンネル ▼ 公募案内・支援 ▼ TOP^ ▼

TOP > NICTについて > 理事長ごあいさつ

理事長 あいさつ



理事長 宮原 秀夫

情報通信研究機構(NICT:National Institute of Information and Communications Technology)は、来るべきユビキタスネット社会を支える情報通信技術の研究開発を、基礎から応用まで一貫した統合的な視点で行い、併せて情報通信分野の事業支援等を総合的に行う独立行政法人です。

私たちは、平成18年4月より、新しい5か年の中期計画期間をスタートしました。この大きな節目を迎えるに当たり、これまで取り組んできた研究開発内容を、「新世代ネットワーク構築技術」、「ユニバーサルコミュニケーション基盤技術」、「安心・安全のためのICT」の3つの研究領域に集約するとともに、これらの研究開発を推進する研究組織についても大幅な見直しを行いました。

情報通信技術(ICT)は、全ての産業活動を支える基盤となる技術分野です。私たちは、ユニバーサルコミュニケーションという概念を掲げ、世界中の人達相互間で自由にコミュニケーションができるようにするとともに、人と機械、機械相互間においても自由にやりとりができる理想の社会の実現を目指して努力しております。

私たちの研究開発成果は国際的な標準化や産業界への技術移転につないでゆきます。また、技術の実用化に向けた産学結集型研究開発や、大学・企業等への研究委託、技術の事業化を加速するためのベンチャー支援やインフラの高度化支援など、幅広い活動も行ってまいります。

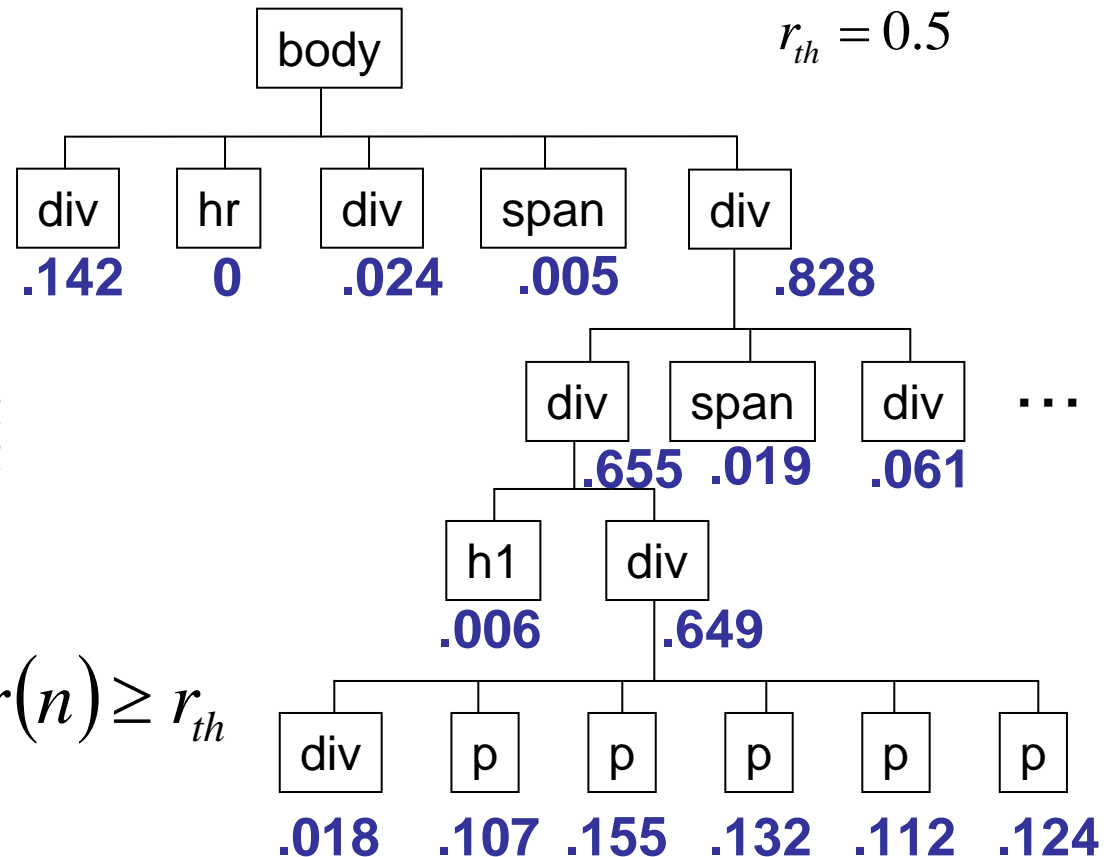
このように、NICTは、情報通信分野における国の唯一の研究機関として、国の情報通信政策を技術的側面から支えるとともに、大学や産業界、さらには海外の研究機関と密接に連携し、また研究成果の社会への普及に積極的に取り組むことにより、活力ある社会、豊かな生活の実現に向けて努力していきます。

Author Name

Main Content

Hypothesis: Author names appear in the proximity of the main content of the page

Recognition of Main Content based on Text Volume



$v(n)$ The volume of text that the sub-tree rooted at node n contains

$$r(n) = v(n)/v(\text{root})$$

$$\arg \min_n r(n) \text{ such that } r(n) \geq r_{th}$$

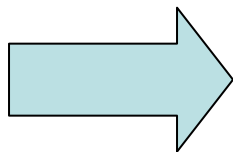
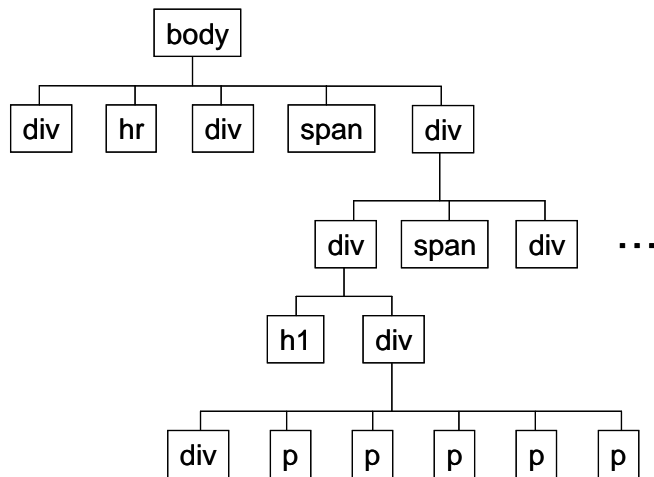
Distance of Elements in a Web Page

- *Rendered Distance*: Distance based on positions of elements after rendering
 - Pros:
 - Reflects the distance that users actually see
 - Cons
 - Rendering is computationally expensive
 - Needs style sheets and image files for accurate rendering
- *Document Structure Distance*: Distance based on document structure
 - Pros:
 - Computationally cheaper than rendering
 - Does not require style sheet or image files
 - Cons:
 - Does not necessarily reflect the distance that the user actually sees

Flattening of DOM Tree

- Purpose
 - Want to measure distance of elements in DOM
 - Want it to be as close to the rendered distance
- A straightforward measure
 - Path length between the nodes in the DOM tree
 - Problem:
DOM trees tend to have deeply nested structure and does not necessarily reflect the rendered distance
- Flattening of DOM
 - Convert a DOM tree into a sequence of block-level elements that directly has text nodes as its children

Example of DOM flattening



Flattening

1	p	NICT 独立行政法人情報通信研究機構
2	span	ナビゲーションをスキップして本文へ
3	form	サイト内検索
4	li	ナビゲーション一覧を開く
5	li	English
	.	
	.	
	.	
17	p	現在位置: ホームの中の機構案内の...
18	h1	理事長あいさつ
19	div	理事長 宮原秀夫
20	p	情報通信研究機構は、来るべきユビキタ スネット社会を支える情報通信技術の...

Distance from the Main Content

d_m Distance from the main content measured on the flattened DOM
(0 when the element is inside the main content)

d_b Distance from the boundary of the main content measured on the flattened DOM

d_c Distance from the center of the main content measured on the flattened DOM normalized by the size of the main content

Distance from the Main Content

			d_m	d_b
1	p	NICT 独立行政法人情報通信研究機構	17	17
2	span	ナビゲーションをスキップして本文へ	16	16
3	form	サイト内検索	15	15
4	li	ナビゲーション一覧を開く	14	14
5	li	English	13	13
	.			
	.			
	.			
17	p	現在位置: ホームの中の機構案内の...	1	1
18	h1	理事長あいさつ	0	0
19	div	理事長 宮原秀夫	0	1
20	p	情報通信研究機構は、来るべきユビキタ スネット社会を支える情報通信技術の...	0	2

Experiment

- Dataset
 - Information Credibility Evaluation dataset [Miyamori et al., 2008]
 - ~2000 pages (20 topics * 100 pages)
 - Each page has ISC annotated by human annotator
 - Used only those pages that has information sender type of `bunch` or `publish` (~400 pages)
- Evaluation
 - Topic-wise 20-fold cross-validation
 - Evaluation measure:
Precision of ranking at cut off rank of 1, 3, and 5
- Results
 - Precision: 58.6% (@1), 72.0 (@3), 75.2 (@5)
(If all the candidates considered: 84.7%)

Successful Example (1) : Government Notice

・戻る

System Output: 北島 新開発食品保健対策室長

平成18年2月13日
厚生労働省医薬食品局食品安全部
北島 新開発食品保健対策室長
担当:調所、柘、岡野(内線2479)

アガリクス(オロシイラクダ)を含む製品の安全性に関する Kitajima, General Manager, Office of Health Policy on Newly Developed Food

- 本朝、アガリクス(オロシイラクダ)を含む製品の安全性に関する食品健康影響評価について、食品衛生委員会に諮問されたことによりお知らせいたします。
1. 厚生労働省では、アガリクスを含む製品が広域に渡って流通していること、アガリクスを含む製品による健康被害が報告されていませんが、肝障害の疑い等複数の事例が学術雑誌等に掲載されていること等から、国立医薬品食品衛生研究所において、アガリクスを含む市販の3製品の毒性試験を実施してきました。
この結果、現在実施しているラットを用いた中期多臓器発がん性試験(※1)において、1製品に発がんプロモーション作用(※2)が認められたとの中間報告を受けました(表1参照)。
 2. これを受け、本日、厚生労働省は食品安全委員会に対し、これらのアガリクスを含む3製品の食品健康影響評価を依頼することとしました。
今回、3製品のうちの1製品については食品としての販売を暫定的に禁止することについて、他の2製品については製品の安全性について念のため評価を依頼することとしました。
 3. 今回の試験結果は、ラットにおいて発がんプロモーション作用が認められたというものであり、ヒトに対してただちにがんを引き起こすという結果ではありませんが、ヒトへの健康被害を未然に防ぐため、厚生労働省では、次のような対応を図ることとしています。
(1)ラットへの発がんプロモーション作用が認められた1製品(キリン細胞壁破砕アガリクス顆粒)の販売者(キリンウェルファーズ(株))に対し、自主的な販売停止と回収を要請
(2)消費者に対し当該製品の摂取を控えるよう、注意喚起する(別紙)こととし、アガリクスに関するQ&Aを厚生労働省のホームページに掲載し、適切な情報を提供
(3)各都道府県、関係団体等に対し、周知の協力を要請するための通知を发出
(4)厚生労働省にアガリクスを含む製品に関する相談専用電話を設置
(連絡先電話番号 03-5253-1111 内線4261~4263)
※平成18年2月27日(月)以降の問い合わせ先は、内線2479、4271、4272
 4. なお、他の2製品のうち、1製品(「仙生露顆粒ゴールド」(販売者:(株)サンドリー(※4)))については、遺伝毒性試験は陰性で、発がんプロモーション試験は実施中であり、腫瘍性病変の増加は確認されていません(※5)。他の1製品(「アガリクスK₂ABPC顆粒」(販売者:(株)サンヘルス))については、遺伝毒性が陰性で、ラットにおける発がんプロモーション試験でも現在腫瘍性病変の増加は認められていないとの報告を受けていますが、最終結果は出ておりません。
これら2製品については、試験結果が出次第、食品安全委員会にその内容を報告するとともに公表することとしています。

Successful Example (2) : Newspaper Article

YOMIURI ONLINE | 読売新聞

毎月1日、15日の2回更新
人気のホームおすすめ情報

認知症フォーラム
全国5会場で開催!

医療と介護

ホーム > 医療と介護 > ニュース

[解説]「アガリクス」で発がん促進

健康食品の安全情報を消費者に

厚生労働省は、キノコ的一种「アガリクス」の健康食品について、一部の製品に発がん促進作用が確認されたと発表した。(医療情報部 中島久美子)

アガリクスは「抗がん効果がある」として、粉末やカプセルの加工品が数多く出回り、がん患者の4人に1人が使っているとの調査もある。結果、実験が終了した1製品に、発がん促進を未然に防ぐため、製品の自主回収を要請、

アガリクスに限らず、健康増進や抗がん効果をうたった健康食品は多いが、人間に使って有効性を調べた信頼できるデータは乏しい。発がんがあやふやなまま使

調べる臨床試験が必要ないためだ。安全性データの提出義務や、食中毒以外では健康被害を疑われる例があっても国に報告する義務もない。

だが、多くの健康食品は特定の成分が凝縮されており、過剰摂取で健康被害が生じたり、通常の食材では微量しか含まれない毒性物質を大量に摂取したりする恐れもある。

そこで同省は昨年、錠剤やカプセル状の食品について、安全性に関するメーカーの自主点検指針を作成した。研究データ収集のほか、日本で食習慣のない成分を使ったり、特定の成分を濃縮したりする場合、毒性試験を行うとしている。ただ、メーカーの自主性に委ねられ、実際にどれほど実施されているか不明だ。

アガリクスの場合、製品を使ったがん患者らが肝臓障害で死亡した例や、発がん作用を示す動物実験結果も過去の学会で報告されているが、こうした健康食品のマイナス情報は利用者に伝わりにくい。

有効性や安全性を客観的に検証する第三者機関や、その情報を消費者にわかりやすく伝える仕組みが求められる。

(2006年2月22日 読売新聞)

System Output: 医療情報部 中島久美子


Kumiko Nakajima, Dept. of Medical Information

Successful Example (3) : Article from a Web site

[All About](#) > [健康・医療](#) > [サプリメント・健康食品](#) > 今までに紹介したダイエットサプリ、すべて見せます！ 2005年、ダイエット宣言！！

[クリップする](#)
[RSS](#)
[メールマガジン](#)

- サプリメント・健康食品
- 利用目的別サプリメント
- ダイエット


サプリメント・健康食品
 ガイド: [マリー 秋沢](#) [取材依頼](#) [問合せ](#)
 毎日を生き活きと元気に過ごすためのサプリ活用法を、分かりやすくガイドが伝授。

System Output: 秋沢

- 症状別
- ボディメイキング
- その他の利用目的
- サプリメント・健康食品配合成分
- コエンザイムQ10
- ビタミン・ミネラル系
- アミノ酸系
- ハーブ系
- 抗酸化系
- 機能系
- サプリメント基礎知識
- 栄養の基礎知識
- サプリメント最新情報

[イムサプリメントとは](#)
[ダイエット食品でカロリーオフ！](#)
[海のカ、フコイダンの魅力とは](#)
[スポンサード・リンク](#)

ダイエット特集 掲載日: 2005年 01月 19日

今までに紹介したダイエットサプリ、すべて見せます！ 2005年、ダイエット宣言！！

この記事の関連タグ:




[携帯に送る](#)
[クリップする](#)
[印刷する](#)
[記事一覧](#)

今年こそは“ダイエットを成功させる！”という方へ
 年末年始はたくさん食べるわりに、運動量が少なく「最近、顔がまっちゃんりしてきた・・・」とお悩みの方はぎっと少なくないはず。
 そこで、「体重をどうにか減らしたい！」とダイエット計画を立てている方に参考にしていただきたいのが、以下の表にあるサプリメントです！



2005年アスリムなボディに！


All About おすすめ情報

- 
女性のお悩みはここで解決！
 月経痛、冷え症…。女性特有の体調不良に効果的なモノとは？
- 
23年間のロングセラー！
 肌質や洗い上がりのタイプで選べる、4種類の天然酵素の洗顔料
- 
疲れと冷えのもとを断つ！
 夏バテ予防にはスパイスを！元気になるレシピその他を紹介！
- 
体脂肪ケアのコーヒーって？
 脂肪の吸収を抑える「コーヒー豆マンノオリゴ糖」を配合！

今月の健康・医療関連おすすめ記事

- [ED治療薬に習慣性はないの？](#)

All About スタイルストア

- 
理想のサプリメントづくり手ブログ
 天然由来の素材だけでつくることはとても難しいんです。
- 研究の結果 コレができました
- [ビタミンとミネラル基礎サプリメント](#)
- [サプリメントにまつわるお役立ちのサプリ](#)

Failed Example (1) : The author is a newswire (not a person)

News section of a portal site

The screenshot shows a news portal page with a navigation bar at the top. The main content area features a news article titled "アガリクス裏づけなき大ブーム" (Agaricus boom without backing). Below the article, there is a search bar and a list of related news items. A search result for "J-CASTニュース" is highlighted with a blue box and the text "Answer: J-CASTニュース". At the bottom of the page, there is a section for "注目情報" (Attention Information) with various advertisements and a small profile picture of a man.

Answer: J-CASTニュース

System Output: 平松

Failed Example (2) : More than one person name near the main content

アガリクス専門店・からだにEショップ

0120-888-690

HOME > よくある質問 | 会員価格詳細 | 御注文方法 | お客様情報 | 問い合わせ | お店紹介

アガリクス商品一覧

アガリクス製品に関わる厚生労働省発表について

お試しバック

アガリクス顆粒品

パウダー抽出品

乾燥アガリクス茸

おすすめ商品

熟成プロポリス

オーエムエックス

発表記事

厚生省発表2/13

厚生省発表3/20

解説

管理者Blog

安全なアガリクスを

当店おすすめの大変では、アガリクスの機能性研究を絶えず行っています。 <メーカーHP>>

大塚アガリクスの抗腫瘍効果研究報告_2006/3/27発表 >>

近畿大農産・応生化、近畿大食薬、大塚共同研究

アガリクス茸の抗酸化成分「AOSA」の発見

[大塚、大阪市立大学名誉教授 三崎旭先生との共同研究]

System Output: 大阪市立大学名誉教授 三崎旭先生

Akira Misaki, Professor Emeritus, Osaka City Univ.

...

今までに、アガリクス...
いるケースが多々見られ、今回の調査で、国内での栽培法の安全性が確認されていること
ております。

弊社でご紹介している(株)大塚のアガリクス製品は、国内にてアガリクス栽培、加工技術を
独自に開発し、12年以上になりますが、一度もクレーム、被害の報告はなく、弊社にてご紹介
を始めて10年の間、同じように健康被害・クレーム等の報告はございません。

厳しく管理された専用施設での栽培にて、信頼できる安全、安心なアガリクス製品を生み出
しています。 <大塚紹介>

キムラ・テック有限会社 代表 木村 正人

Answer: 木村正人 (Masato Kimura)

[アガリクス専門店] [熟成プロポリス] [純生ローヤルゼリー] [オーエムエックス] [黒糖花粉]
[梅肉エキス] [乾燥納豆] [森の薬] [はちみつ黒糖バーモン] [黒糖王(黒糖もちろみ)] [竹炭豆]
[マカ+CoQ10] [α-リポ酸+ヒアルロン酸] [エソウキ新芽茶] [ギャバ煎茶] [ペット用アガリクス]
[コンピュータ用伝票] [蛍光灯用反射板Light Up] [KIM-TEC@Users] [寝たつき餅神田] [WEB健康(倶楽部)]

キムラ・テック(有)通販事業部
広島県福山市手塚町2-2-4 TEL: 0120-888-690

30 Oct 2008


Failed Example (3) : Blog

漢方で1日1善

あなたにも当てはまるケースがありませんか？

当店に来られる漢方相談の中から症例と漢方薬処方方をできるだけ毎日記録していくようにいたします。
ずっと一人で悩んでいたこと、知りたかったこと・・・ここで見つけてください。

ICHIZEN



古村学 先生

プロフィール

最近のコメント

- いちべし on 夜間頻尿
- 鈴木佳代 on 夜間頻尿
- いちべえ on 飛蚊症の漢方
- 高木央子 on 飛蚊症の漢方
- いちべえ on 水イボ 1
- 三児のママ on 水イボ 1
- ろんど on 飲食不可能な方に

◀ [皮膚掻痒症](#) | [トップページ](#) | [田七の効きめ](#) ▶

2006年2月14日 (火)

アガリクス安全性について

Answer: 古村学

アガリクス製品の発ガンプロモーション作用が認められたとの報道がありました。この中では、3社の製品を対象に試験が行われ、そのうちの1社に問題があったとのことで、他の2社は発がん性はないが、その他の継続試験中のようです。

System Output: 鈴木佳代

アガリクスは日本でも使用の歴史は古いうで、多分30年弱使われてきたと思います。当店も最近では使用量が少なくなりましたが、現在はブラジル産の乾燥品のみ扱っています。しかし最近では健康ブームで様々なアガリクス製品が普及してしまいましたが、薬効を考えるとやはり煎じ薬がお勧めです。加工することによって生薬の効果が変化するだけでなく、各種添加物も使う必要が発生します。

今回のアガリクスの問題は、その原因がどこにあるか容易には突き止められないと思いますが、考えられることはアガリクスそのものの問題より、

最近の記事

- [高齢者の残尿・頻尿トラブル](#)
- [様々な病名と漢方](#)
- [慢性荨麻疹が改善](#)
- [薄くなる髪](#)
- [漢方薬の指定買い](#)
- [料理屋さんの職業病](#)
- [中国のダイエット医薬品に注意](#)
- [中国の冬虫夏草](#)
- [成都案内2](#)
- [成都案内1](#)

2006年11月

日	月	火	水	木	金	土
			1	2	3	4
5	6	7	8	9	10	11

Learned Model: Positive Features

Summed weights

Features

486.975525

person

343.295375

context_tag_font

314.868000

num_morph

303.745805

context_tag_select

303.745805

context_tag_option

270.825498

context_tag_b

154.405898

WORD_IPS

137.443534

location

131.968790

startword_IPS

124.140339

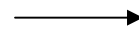
context_word_日

118.839416

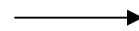
startword_松崎

118.839416

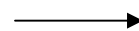
WORD_松崎



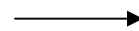
Features based on linguistic analysis



Features based on document structure



Context of the author name
日 (Day) is used for date expressions



Tokens that are specific to certain instances

Learned Model: Negative Features

Summed weights	Features	
-189.670375	context_tag_span	→ Features based on document structure
-132.047854	context_tag_img	
-104.351307	distance_from_main_edge	Distance based feature
-38.114124	tag_meta_keywords	
-38.060345	context_word_方	→ Relatively general context words
-33.739126	WORD_ダイエット	
-31.354978	WORD_会	→ Topic specific tokens
-30.968525	context_word_写真	
-29.672442	startword_ダイエット	→ Relatively general tokens
-28.682960	startword_日本	
-28.491197	context_word_サイト	
-27.058144	WORD_市	

Weights of other distance features: -0.673151 distance_from_main
-19.485953 distance_from_main_center

Observations

- The method works when the author name appears near the main content (*this is what we have expected*)
- Performs poorly when:
 1. the author name do not appear near the main content (as in blogs)
 2. there are more than one person names near the main content
 3. the author is not a person but an organization

Possible Improvements

- Add more features:
 - which can be used to identify author names that are relatively far from the main content
 - which can distinguish the author name from other person names which are not the author
- More data:
 - There may be not enough instances where the author is an organization for the model to learn
30 / 400 ~ 7.5% (rest are individuals)
- Web page structure recognition
 - Recognize the semantics of the parts of a Web page
 - Main article, profile, biography, link list, ads, etc.

Summary

- Approach to information credibility analysis from the aspect of *information sender*
- Description of information senders and their relationships in terms of *information sender configuration (ISC)*
- Extraction of the author of Web pages as a sub-problem of ISC identification
- Evaluation results: Precision at ~60% (75% @3)
- Challenges ahead:
 - Improving precision
 - Dealing with multiple authors (forum posts, product review comments)
 - Identifying parts of a Web page that come from different authors, and their corresponding authors
 - Expertise analysis

Merci beaucoup

ykato@nict.go.jp