

検索エンジンの仕組み

東京大学

イノベーション政策研究センター

ウェブ World Wide Web

- 1990 最初のウェブページ (ティムバーナーズリー、CERN)
- インターネット上のドキュメント(ウェブページ)をハイパーリンクにより相互に参照可能にする仕組み



IPDLへのハイパーリンク
特許公報などの検索(IPDL)



検索エンジン

- ウェブページを検索するための仕組み
- 検索エンジンを提供するサイト
 - Google
 - Yahoo
 - Bing (Microsoft)
- 国内
 - goo (NTT)
 - Infoseek (楽天)
 - Excite

すべての結果

[画像](#)[動画](#)[ニュース](#)[ブログ](#)[アップデート](#)[書籍](#)[掲示板](#)

期間指定なし

[最新](#)[24 時間以内](#)[1 週間以内](#)[1 年以内](#)[特定の期間](#)

すべての結果

[閲覧したページ](#)[未閲覧のページ](#)

標準ビュー

[ワンダーホイール](#)[タイムライン](#)

基本情報のみ

[画像表示](#)[プレビュー表示](#)[翻訳して検索](#)他のキーワード: [特許庁 電子図書館 検索](#) [特許庁 商標](#) [特許庁 特許電子図書館](#) [特許庁 データベース](#) [特許庁 ホームページ](#)[特許庁ホームページ](#) ☆

知的創造活動によって生み出されたもの(アイデア、デザインなど)を、創作した人の財産として保護する。

www.jpo.go.jp/indexj.htm - [キャッシュ](#) - [類似ページ](#)[弁理士試験](#)[出願受付](#)[商標について](#)[初めての方](#)[特許について](#)[手続が分からない時は](#)[採用情報](#)[無料で先行技術が調査できる](#)jpo.go.jp からの検索結果 »[特許庁ホームページ - Japan Patent Office](#) ☆

三角 近年の法律・政令・省令改正については、「特許庁の取り組み」の「法令改正のお知らせ」に掲載しています。 ... 出願等統計（2007年版）について（特許行政年次報告書2007年版「統計・資料編」へ） 2007.6.19 ...

www.jpo.go.jp/shiryou/index.htm - [キャッシュ](#) - [類似ページ](#)[特許庁 - Wikipedia](#) ☆

特許庁（とっきょちょう、Japan Patent Office）は、日本の中央省庁の一つ。経済産業省の外局。産業財産権（工業所有権）制度の企画立案と審査・審判などを所掌する。初代長官（農商務省専売特許局長、及び特許局長）は高橋是清。 ...

ja.wikipedia.org/wiki/特許庁 - [キャッシュ](#) - [類似ページ](#)[\[INPIT\] 特許電子図書館（IPDL）サービス一覧 | 独立行政法人 工業...](#) ☆

人と情報で知財を支えるINPIT―「工業所有権情報・研修館」では、特許庁と連携しつつ、特許情報提供、特許流通促進、... 特許電子図書館（IPDL）サービス一覧内の各サービスは、特許電子図書館（IPDL）サイト内の各サービスにリンクしております。 ...

www.inpit.go.jp/ipdl/service/index.html - [キャッシュ](#)[특허청](#) ☆ - [[このページを訳す](#)]

특허청 소개, 산업재산권제도 및 절차안내, 산업재산권정보 이용안내, 민원안내, 특허정보 검색.

www.kipo.go.kr/ - [類似ページ](#)[特許庁・商標検索サービス](#) ☆

特許庁商標検索サービスへリンクします。 ... 実際に登録するには既に商標として登録されていないか、また商標出願されていないかを特許庁・無料商標検索サービスで検索してみます。検索する類似群コードが不明な場合は商品・役務名リスト（特許庁）で ...

www.psn.ne.jp/~bds/kentokyu/tokyochou/kensaku.htm - [キャッシュ](#) - [類似ページ](#)[United States Patent and Trademark Office](#) ☆ - [[このページを訳す](#)]

About USPTO · Executive Biographies · USPTO Offices · Budget, Performance, and Planning · Statistics · Vendor Information · Public Advisory Committees · National Medal of Technology and Innovation · Director's Forum · Feedback Channel

検索結果

検索エンジンの仕組み

- ウェブページをどのように集めるか
 - ← クローリング
- 集めたウェブページをどのように管理するか
 - ← インデキシング
- 検索クエリーとウェブページをどのようにマッチングするか
 - ← ランキング

検索エンジンの仕組み



クロール

インデックス

ランキング



検索エンジン

- ロボット型 (1998～)
 - ウェブページをプログラムにより自動でクローリング、インデキシング、ランキング
- ディレクトリ型 (～1997)
 - ウェブページを人手で収集、分類
質の高いページを集める事ができるが検索対象
ページの数を増やす事が難しい
 - Googleは数十億のページを検索対象

検索エンジンの仕組み

クローリング

- クローラー(スパイダー)と呼ばれるプログラムによりウェブページのハイパーリンクを自動で辿っていき、ウェブページを収集
- 収集したウェブページはインデキシングをして管理する
- Googleのクローラ
 - 66.249.68.121 - - [13/Apr/2010:04:39:02 +0900]
"GET /index.html HTTP/1.1" 200 36232 "-"
"Mozilla/5.0 (compatible; **Googlebot**/2.1; +http://www.google.com/bot.html)"

検索エンジンの仕組み インデキシング

- ウェブページからキーワード(索引語)を抽出
- キーワードの重みを計算
- キーワードとウェブページの転置索引を作成



	特許庁	INPIT	IPDL
特許	10	3	5
知的財産	4	8	2
図書館	3	5	10
審査請求	8	5	8

検索エンジンの仕組み ランキング

- クエリーとウェブページの関連性を計算
- 転置索引を参照

特許

☒ ウェブ全体から検索 ☐ 日本語のページを検索



	特許庁	INPIT	IPDL
特許	10	6	7
知的財産	2	8	3
図書館	3	5	10
審査請求	8	5	8

- 1** [特許庁ホームページ](#) ☆ - 2 回閲覧 - 22:03
知的創造活動によって生み出されたもの(アイデア、デサ
て保護する。
[弁理士試験](#) - [商標について](#) - [特許について](#) - [採用情報](#)
www.jpo.go.jp/indexj.htm - [キャッシュ](#) - [類似ページ](#)
- 2** [\[INPIT\] 特許電子図書館 \(IPDL\) サービス一覧 | 独立行政法人工業...](#) ☆
特許電子図書館 (IPDL) サービス一覧内の各サービスは、特許電子図書館 (IPDL) サイト内
の各サービスにリンクしております。特許電子図書館 (IPDL) トップページ, 「特許電子図
書館 (IPDL) サービス一覧」に関するページ, 特許電子図書館 (IPDL) ...
www.inpit.go.jp/ipdl/service/index.html - [キャッシュ](#)
- 3** [\[INPIT\] 独立行政法人 工業所有権情報・研修館](#) ☆
人と情報で知財を支えるINPIT-「工業所有権情報・研修館」では、!
許情報提供、特許流通促進、産業財産権相談、知財人材育成などの
www.inpit.go.jp/ - [キャッシュ](#) - [類似ページ](#)

検索エンジンの仕組み ランキング

- ANDやORのクエリー検索も同様
- 転置索引を参照



	特許庁	INPIT	IPDL
特許	10	6	7
知的財産	2	8	3
図書館	3	5	10
審査請求	8	5	8

検索

特許 知的財産

☒ ウェブ全体から検索 ☐ 日本語のページを検索

1

[\[INPIT\] 独立行政法人 工業所有権情報・研修館 ☆](#)
人と情報で知財を支えるINPIT-「工業所有権情報・研修館」では、特許情報提供、特許流通促進、産業財産権相談、知財人材育成などの業務を
www.inpit.go.jp/ - [キャッシュ](#) - [類似ページ](#)

2

[特許庁ホームページ ☆](#) - 2 回閲覧 - 22:03
知的創造活動によって生み出されたもの(アイデア、デザイン)を保護する。
[弁理士試験](#) - [商標について](#) - [特許について](#) - [採用情報](#)
www.jpo.go.jp/indexj.htm - [キャッシュ](#) - [類似ページ](#)

3

[\[INPIT\] 特許電子図書館 \(IPDL\) サービス一覧 | 独立行政法人 工業所有権情報・研修館 ☆](#)
特許電子図書館 (IPDL) サービス一覧内の各サービスは、特許電子図書館 (IPDL) サイト内の各サービスにリンクしております。特許電子図書館 (IPDL) トップページ、「特許電子図書館 (IPDL) サービス一覧」に関するページ. 特許電子図書館 (IPDL) ...
www.inpit.go.jp/ipdl/service/index.html - [キャッシュ](#)

検索エンジンの仕組み

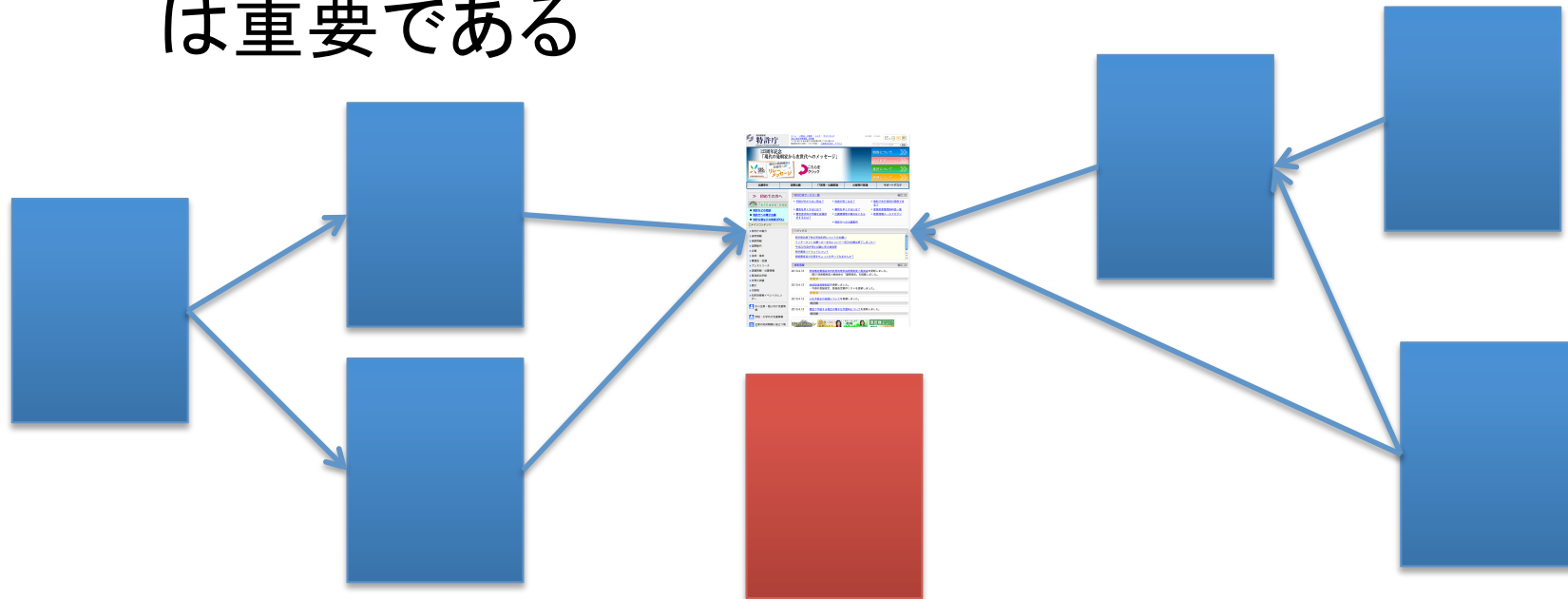
ランキング

- クエリーとウェブページの関連度だけではスパムページ(悪意あるページ)が上位来てしまう可能性
 - 悪意あるユーザが作成した“特許”というキーワードをたくさん含んだ特許とはまったく関係のないページ

→ ウェブページの重要度を計算

検索エンジンの仕組み ランキング

- ウェブページの重要度を計算
- PageRank (P. Larry and S. Brin, 1998)
 - Googleのランキング技術
 - 多くの重要なページからリンクされているページは重要である



検索エンジン データセンタ

- 数十億のウェブページのクロール、インデキシング、ランキング

→

- 大規模分散システムが必要
 - 数十万台のサーバを世界中に配置
 - クラウドコンピューティング
 - Google File System 分散ファイルシステム
 - MapReduce 分散処理フレームワーク
 - Big Table 分散データベース

検索エンジン最適化 (SEO)

- ランキングは一般に検索エンジン企業の秘密
- 検索エンジンでページが上位にくるには
 - クエリーに関連したキーワードを増やす
 - キーワードをページの重要な部分に置く
 - ハイパーリンクを張る、もらう
- 過剰なSEOは検索エンジンにはじかれる

まとめ

- ウェブ
 - ハイパーリンクで結ばれたドキュメント(ウェブページ)
- 検索エンジン
 - クローリング
 - プログラムによるウェブページの自動収集
 - インデキシング
 - 収集したウェブページにキーワード(索引)付け
 - ランキング
 - クエリーとキーワードのマッチング
 - PageRank

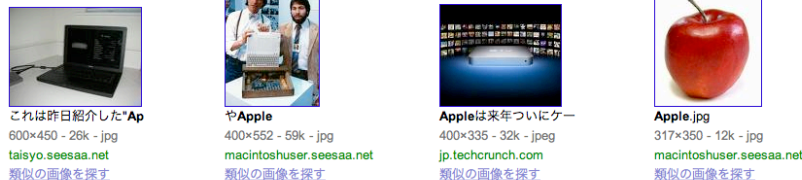
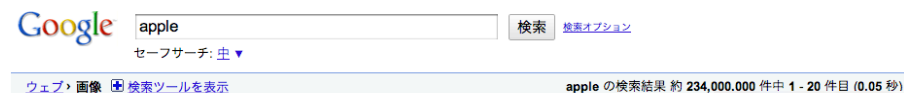
ウェブ検索と意匠・商標

東京大学

イノベーション政策研究センター

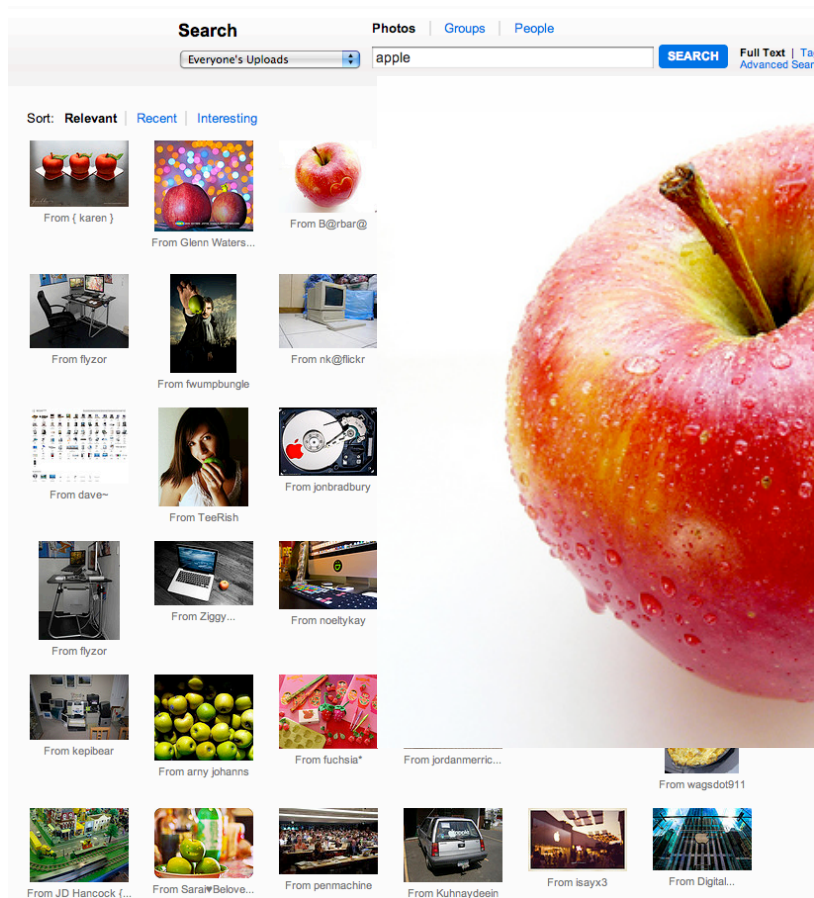
ウェブ検索と意匠

- 画像検索
－ 画像の周囲のテキストを検索



ウェブ検索と意匠

- 画像検索
 - 人手でつけたタグを検索



Tags

- cuore
- heart
- mela
- apple
- frutta
- fruit
- rosso
- red
- amore
- love
- goccia
- gocce
- drop
- drops
- ArteDellaFoto
- beautiful secrets

ウェブ検索と意匠

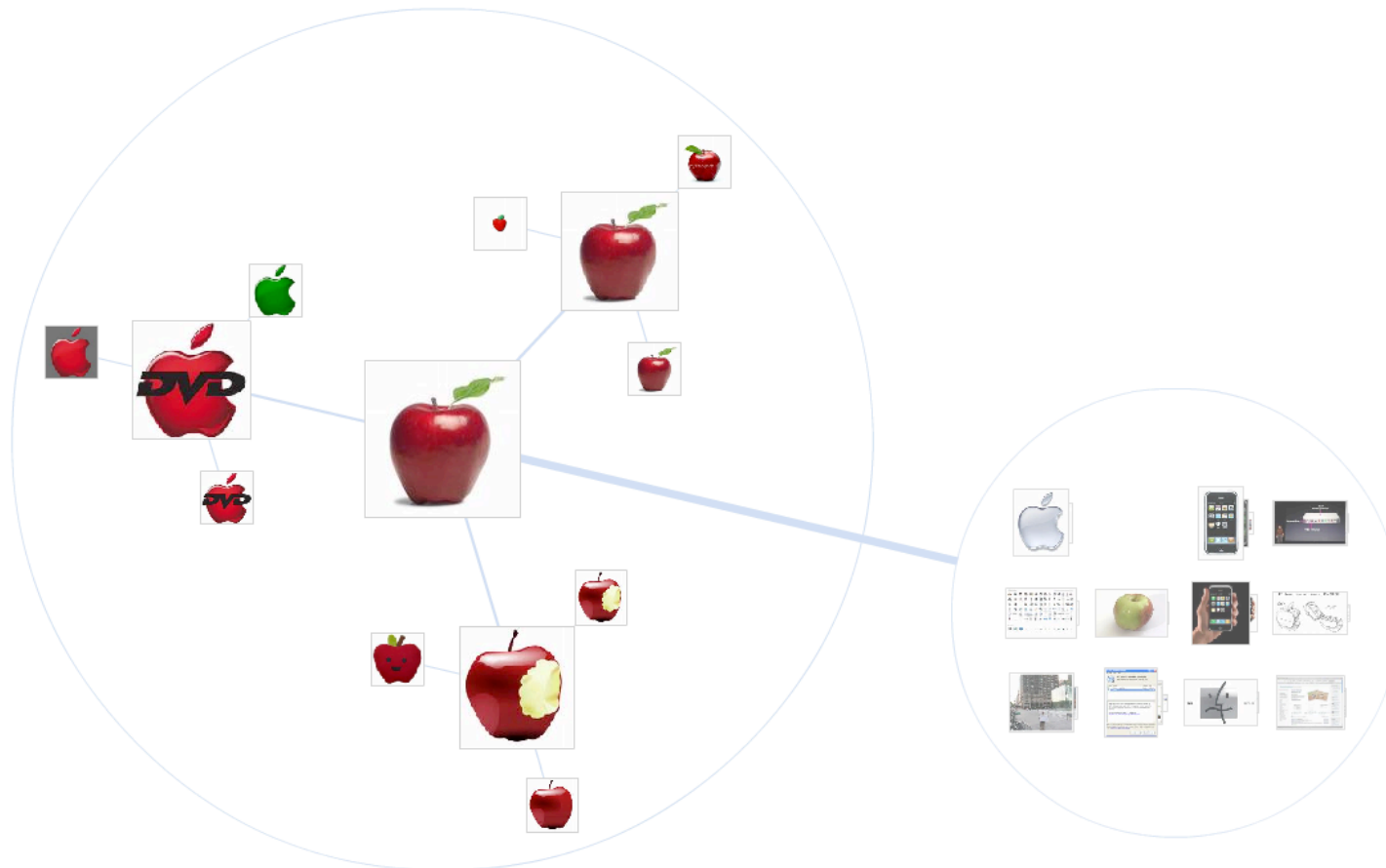
- 画像周りのテキストよりもタグは正確
- タグを増やすには?



ウェブ検索と意匠

- テキスト/タグでなく画像そのものの検索
- 類似画像検索





ウェブ検索と商標

- 類似文字列検索
 - 似ている言葉を探す自然言語処理技術



スパゲッティー:

スパゲッティ スパゲッテ スパゲティ スパッティ スパゲッティ スパゲッ
ティニ スパゲッティ・ スパッゲッティ スパゲッティニ・・・スパゲッテ
ィー スーブスパゲッティ スパゲッティmonster

ウェブ検索と商標

- 信憑性(Trust)の高い情報のウェブ検索
 - 評判情報抽出
 - 実世界情報とのマッチング

jpo.go.jp



人気度：
順位：
オーナー：
サーバーの場所：
統計データ：

説明：
JPO:Japan Patent Office

該当データなし
whois
JP
Compete

Congratulations on your site's reputation

Increase sales, enhance competitiveness & boost customer trust on your website with the WOT trust seal. Try it free for 30 days

サイトの格付け：

信頼性	90	
業者の信頼度	94	
プライバシー	98	
子供の安全	99	

カテゴリ別コメント：

コメントの追加

コメントのリクエスト

Comments provide more information, but do not affect ratings.

日付	ソース	カテゴリ	コメント
----	-----	------	------

現在の検索から セマンティック技術へ

東京大学

松尾 豊

森 純一郎

検索エンジン

- クローリング
- インデキシング(索引付け)
- 結果の提示
 - PageRank
 - 文書とクエリーのマッチ度
 - SPAM判定(例えばコンテンツの重複など)
- 関連性フィードバック
 - Learning to Rank

キーワードの検索から 意味の検索へ

- 東京大学⇒東大、University of Tokyo
 - クエリーの共起や翻訳
- 東京大学 is-a 大学機関
- 東京大学、京都大学
 - 大阪大学、慶應大学、早稲田大学、・・・大学
- 東京大学の学長は？

東京大学

名称	東京大学
学長	濱田純一
所在地	東京都文京区
学生数	14000人(学部)
予算	

京都大学

名称	京都大学
学長	松本紘
所在地	京都府京都市
学生数	
予算	

大阪大学

名称	大阪大学
学長	
所在地	
学生数	
予算	

慶應大学

名称	慶應義塾大学
学長	
所在地	
学生数	
予算	

...

東京大学

名称	東京大学
学長	濱田純一
所在地	東京都文京区
学生数	14000人(学部)
予算	

京都大学

名称	京都大学
学長	松本紘
所在地	京都府京都市
学生数	
予算	

濱田純一

名前	濱田純一
生年月日	1950年3月14日
出身大学	東京大学
学外における役職	日本マス・コミュニケーション学会会長
学外における役職	総務省電波監理審議会会長

日本マスコミュニケーション学会

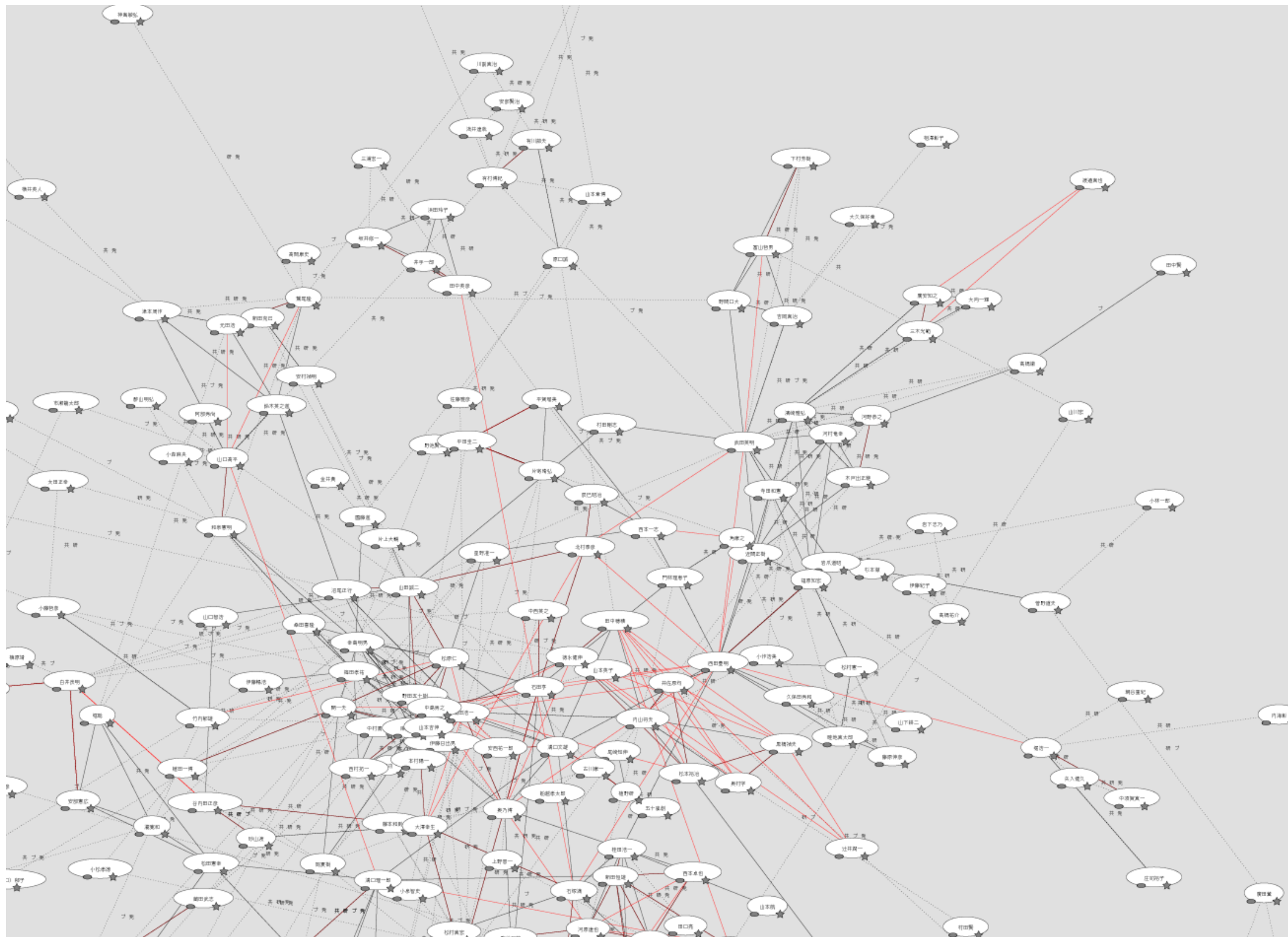
名称	日本マスコミュニケーション学会
学長	濱田純一
所在地	千代田区三崎町
旧名称	日本新聞学会
会員数	

エンティティと属性、関係性

- 人名、組織名、地名、商品名、技術名、...
- 生年月日、所在地、出身地、製造元、基盤技術、...
- 「代表を務める」「所属している」「友人」「製造元」「後継商品」...
- 人工知能：
 - 意味ネットワーク (Richard H. Richens, 1956)

次世代の検索

- Microsoft Bing
 - オバマ大統領はどこで生まれましたか？
- 知識獲得のボトルネック
- 大量の情報が利用可能になるにつれて、新しい形でのAIの活用が可能になりつつある。
 - 研究者と専門分野の関係
 - 企業と人の関係
 - 要素技術と製品の関係
- 知識の「断片」を取り出し、組み合わせることで、新たな知識として活用することができる。



Web Two names Results 1 - 10 of about 742 for "Yutaka Matsuo" "Mitsuru Ishizuka". (0.41 seconds)

DBLP: Yutaka Matsuo

www.informatik.uni-trier.de/~ley/db/indices/a-tree/m/Matsuo:Yutaka.html - 19k -

[Cached](#) - [Similar pages](#)

DBLP: Mitsuru Ishizuka

19. Mitsuru Ishizuka, Yutaka Matsuo: SL Method for Computing a ...

www.informatik.uni-trier.de/~ley/db/indices/a-tree/l/lshizuka:Mitsuru.html - 66k -

Cached - Similar pages

Average-clicks

Ishizuka, Graduate School of Information Science and Technology, ...

portal.acm.org/citation.cfm?id=1008494&

dl=ACM&coll=ACM&CFID=15151515&CFTOKEN=6184618 - Similar pages

Yutaka Matsuo

Ishizuka: Finding Social Network for Trust Calculation, Proc. ...

www.ymatsuo.com/publication_eng.htm - 18k - [Cached](#) - [Similar pages](#)

yutaka matsuo - ResearchIndex document query

Cost-based Hypothetical Reasoning Yutaka Matsuo and Mitsuru Ishizuka ...

citeseer.ist.psu.edu/cis?q=Yutaka+Matsuo - 18k - Cached - Similar pages

[Home Page of Ishizuka Laboratory](#)

y.matsuo@carc.aist.go.jp; Naohiro Matsumura (03 Dr.) -- matumura ...

www.miv.t.u-tokyo.ac.jp/HomePageEng.html - 18k - Cached - Similar pages

Try your search again on [Google Book Search](#)

Publication

-> coauthorship

My homepage

Laboratory page

-> same laboratory relation

抽出の方法

ウェブ イメージ グループ ディレクトリ

Google 松尾豊 石塚満 Google 検索 検索オプション 表示設定

ウェブ全体から検索 日本語のページを検索

ウェブ 松尾豊 石塚満 の検索結果 約 124 件中 1 - 10 件目 (0.17 秒)

主要邦文論文 -- 石塚研究室

... 松尾 豊, 石塚 満: 語の共起の統計情報に基づく文書からの
キーワード抽出アルゴリズム, 人工知能学会論文誌, Vol.17, No.3,
pp.217-223 (2002.5) [人工知能学会論文賞受賞] [PDF]. 松尾 豊, 石塚 ...
www.miv.t.u-tokyo.ac.jp/fj-pub.html - 16k - 2004年4月24日 - [キャッシュ](#) - [関連ページ](#)

松尾 豊の研究内容

... 松尾 豊, 石塚 満: コストに基づく仮説推論の2種の連続値最適化問題
への置換法とその協調による推論法, 人工知能学会誌, Vol.16, No.4(B), pp.400-407,
2001 [PDF 61k] 仮説推論の数学的な制約式への置き換え法が2種類あるよ ...
30k - [キャッシュ](#) - [関連ページ](#)

2 題目, 1F1-02 Web

松尾 豊 友部 博教

です。論文, PDFファ

297.HTM - 3k - [キャッシュ](#) - [関連ページ](#)

る。所属,

40 F 小会

IS/PE00182.HTM - 4k - [キャッシュ](#) - [関連ページ](#)

石塚 満

石塚 満 氏名, 石塚 満 ぶりがな, いしづか みつる, 所

小会

- 4k - [キャッシュ](#) - [関連ページ](#)

の抽出と情報支

メ干, Webマイニ

例)124件

検索エンジン(Google)を使って、共起関係の強さを測る。

Jaccard係数、相互情報量などさまざまな尺度があるが、
閾値つきOverlap係数を用いる。

例)

“松尾豊 石塚満”: 124件 強い

“松尾豊 溝口理一郎”: 11件 弱い

“石塚満”: 791件

“溝口理一郎”: 813件

$$R(X,Y) = \begin{cases} \frac{|X \cap Y|}{\min(|X|, |Y|)} & \text{if } |X| > k \text{ and } |Y| > k, \\ 0 & \text{otherwise} \end{cases}$$

検索されたWebページから、ページの特徴量を抽出し関係を把握。

共著、研究室、プロジェクト、発表

Simpson係数は、関係の強さを的確に表す。

9割程度の適合率(再現率は2割~5割): アンケート調査(141名)

関係の判別

“X and Y”で検索した上位ページを対象

属性リスト

- 2人の氏名の共起回数 • Simpson係数が閾値以上か
- Xの出現回数 • Yの出現回数
- {出版、論文、発表、活動、テーマ、賞、著者}のいずれかの語がタイトルに含まれるか
- {メンバー、研究室、研究所、研究機関、チーム}のいずれかの語がタイトルに含まれるか
- {ワークショップ、会議、セミナー、ミーティング、スポンサー、シンポジウム}のいずれかの語がタイトルに含まれるか
-
- {出版、論文、発表、活動、テーマ、賞、著者}のいずれかの語が最初の5行に含まれるか
-

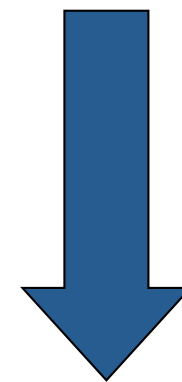
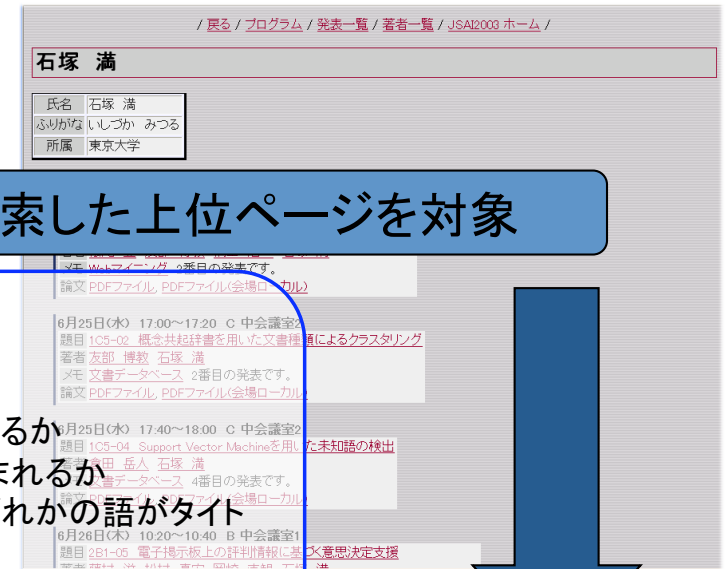
ページの特徴属性

(more than one, yes, yes, more than one, more than one, no, no, no, no, no, no, yes, no, no, no, yes, no)

判別ルール

- NumCo = more_than_one → 共著
- NumCo = more_than_one & GroFFive(F)=no → 研究室
- (Rel=yes & GroTitle(E)=no & GroFFive(C)=no → 研究室
-

関係のクラス: 共著、研究室、プロジェクト、発表



技術的な難しさ

- 同姓同名・別名
 - Mining for Personal Name Aliases on the Web, WWW2008
 - Personal Names on the Web using Automatically Extracted Key Phrases, Proc. 17th European Conference on Artificial Intelligence (ECAI-06)
- 多種のエンティティへの拡張
 - Extracting Social Networks among Various Entities on the Web, Proc. Fourth European Semantic Web Conference (ESWC 2007), June 2007
- 関係抽出: 特定の関係性をどう抽出するか
 - Relation Extraction from Wikipedia Using Subtree Mining, Proc. Twenty-Second Conference on Artificial Intelligence (AAAI-07)
 - Extracting Keyphrases to Represent Relations in Social Networks from Web, Proc. 20th International Joint Conference on Artificial Intelligence (IJCAI2007)
- 学習: どのようにパターンを学習するか
 - Generating Social Network Features for Link-based Classification, Proc. European Conference on Principles and Practice of Knowledge Discovery in Databases, (PKDD07)
 - Measuring Semantic Similarity between Words using Web Search Engines, Proc. 16th International World Wide Web Conference (WWW2007)

POLYPHONET

- 学会支援システム
 - 学会参加者の相互のコミュニケーションを支援する
 - どういう専門家が参加しているのか、誰と研究しているのかを見せる
- 運用実績
 - 人工知能学会で03年から08年まで運用。その他、ワークショップ、国際会議など。

	JSAI03	JSAI04	JSAI05	UbiComp05
#participants	558	639	about 600	about 500
#users	276	257	217	308

JSAI2003

UbiComp05



Polyphonet JSAI2005: Person view

POLYPHONET CONFERENCE

発表検索 検索 研究者検索 検索

マイページ スケジュール つながり検索 ActionLog システム説明 ヘルプ

研究者情報

Yutaka Matsuo

松尾 豊
Yutaka Matsuo

産業技術総合研究所
専門分野 分野2

周りの人達を表示する
研究キーワードを見る

情報の編集

TelMeA Theatre - 新着発言が21件あるよ!

知り合い - すべて表示

双方向 (3人) / 知っている (4人) / 知られている (6人)

松尾 豊 さんの発表 - すべて表示

6/15 16:30- (A会場)
研究者ネットワーク抽出検索システム

6/15 17:00- (C会場)
論文データベースからの研究トピック抽出

スケジュールに追加した発表 - すべて表示

6/16 11:30- (B会場)
Community Webプラットフォーム

6/16 14:00- (C会場)
ソーシャルネットワークとコラボレーション支援

イロノミー検索・推薦

イロノミー検索

☒ 赤 ☒ 青 ☒ 緑 ☒ コメント ☐ 他の人のチェックも検索

イロノミー推薦

周囲の人間関係ネットワーク

[ネットワーク図を大きく表示]

関係の強い人達 - すべて表示

■ 共著 ■ 研究 ■ プロ ■ 発表 ■ 知人

石塚 満
東京大学

大向 一輝
総合研究大学院大学

橋田 浩一
産業技術総合研究所

松村 真宏
大阪大学大学院経済学研究科

武田 英明
国立情報学研究所

濱崎 雅弘
総合研究大学院大学

Research cluster
and Keywords

Acquaintances
(registered by users)

His presentation

Presentation
bookmarks

Retrieval of
annotated comment

Social network

Strong relation
person (Web-mined)

まとめ

- セマンティック技術
 - エンティティと関係性の認識が基本
- 国際的にも盛んに研究されている
 - Google, Yahoo!, Microsoft Researchや米国の大学等
- 「検索」から、文書内にある知識の抽出と、それを用いた分析・編集技術へ。