第3回教員ゼミ

Combining Algorithms Using Metalearning

From “Survey of Text Mining”

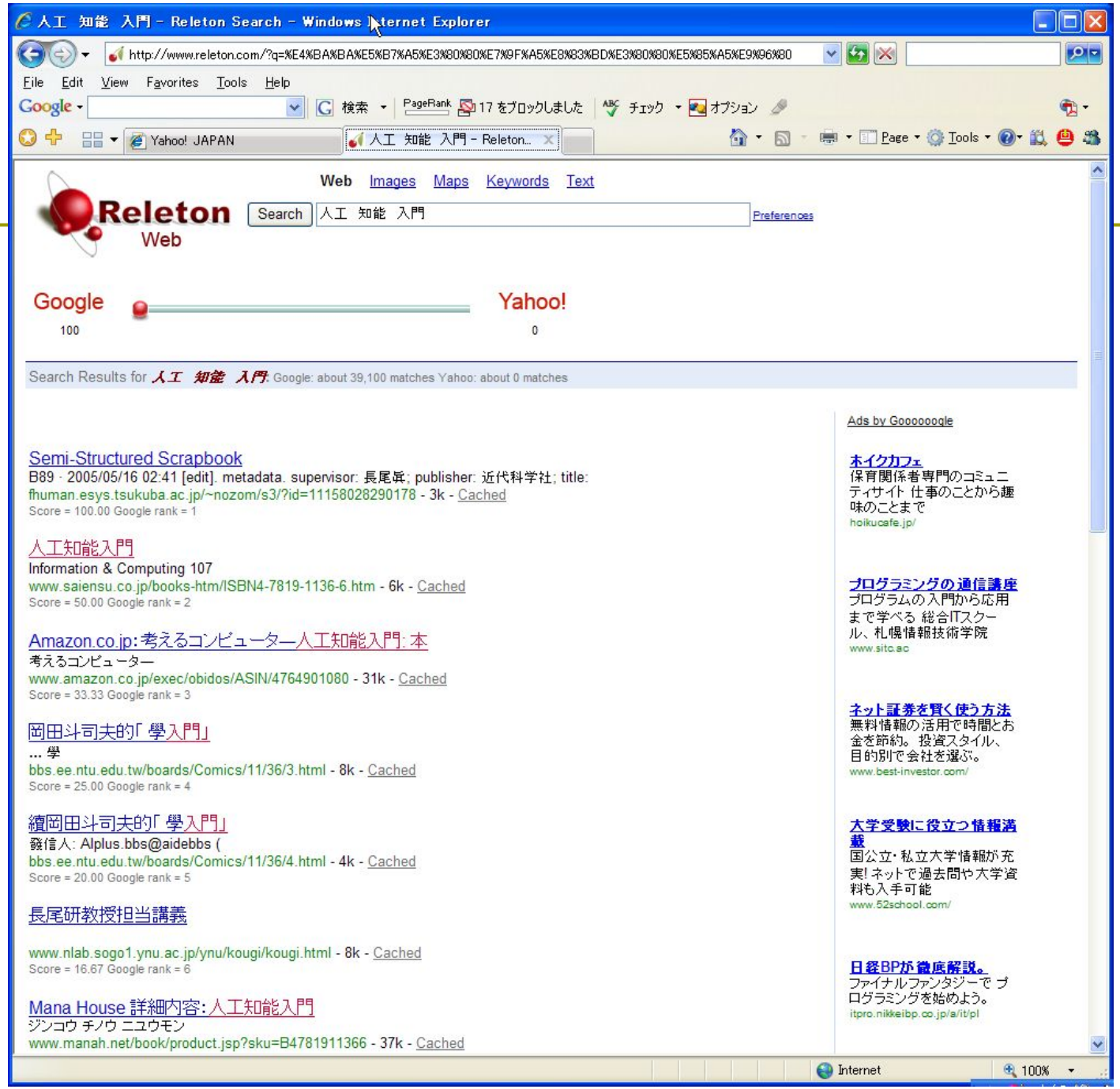
佐々木　稔

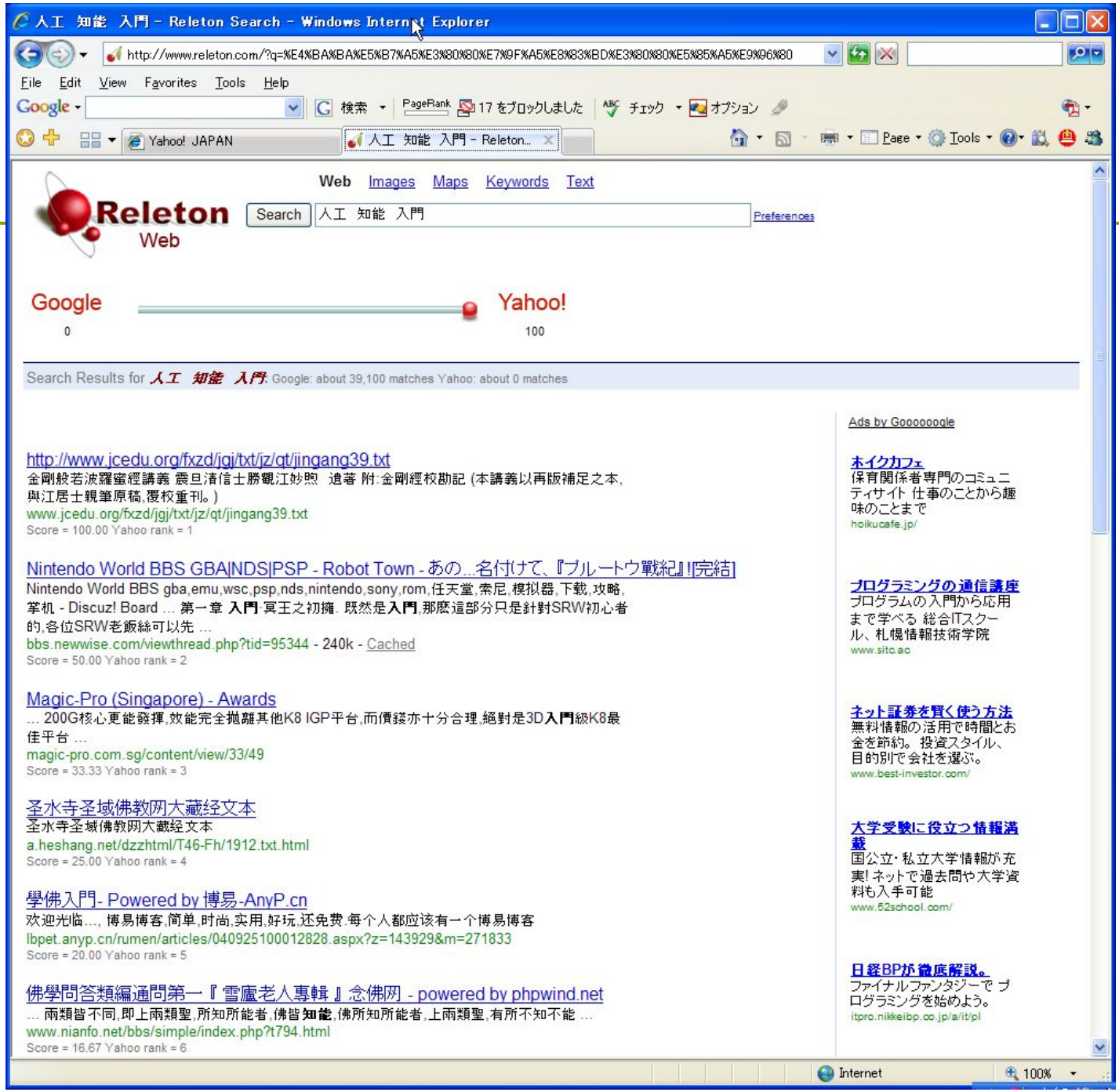
目次

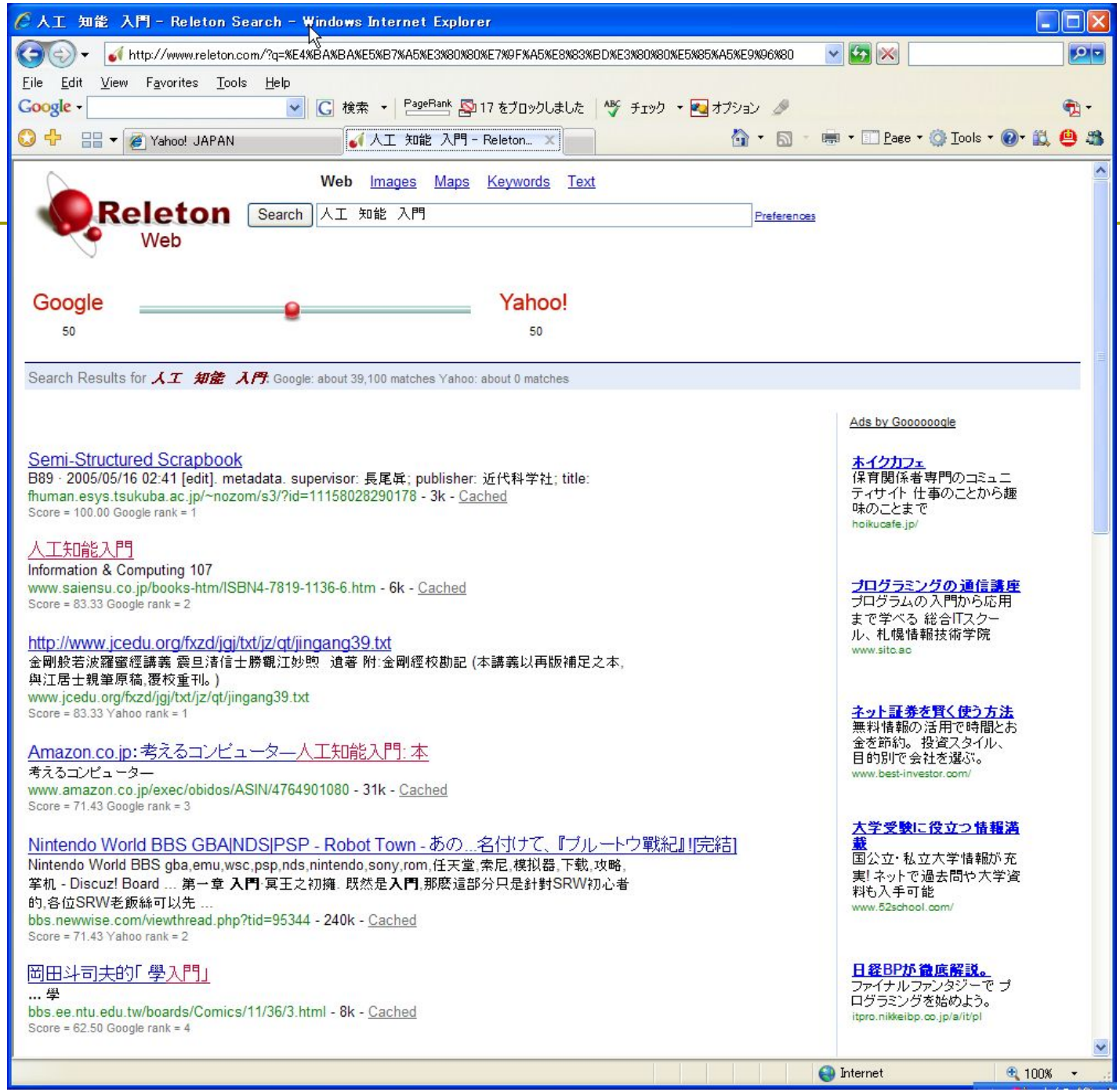
- はじめに
- 関連研究
- ベクトル空間モデル
- Metalearning
- Implementation
- 実験、結果
- 追加実験
- おわりに

はじめに(1/2)

- 検索(IR)エンジン
  - ひとつだけのアルゴリズムは最適ではない
  - IRアルゴリズムの使い分けができないか？
    - ある質問と文書集合の組に適したアルゴリズム
    - 他の質問と文書集合の組には別のアルゴリズム


人工 知能 入門 - Releton Search - Windows Internet Explorer
http://www.releton.com/?q=%E4%BA%BA%E5%B7%A5%E3%80%80%E7%9F%A5%E8%83%BD%E3%80%80%E5%85%A5%E9%96%80
File Edit View Favorites Tools Help
Google 検索 PageRank 17 をブロックしました チェック オプション
Yahoo! JAPAN
人工 知能 入門 - Releton...
Page Tools
Web Images Maps Keywords Text
Releton Web
Search
人工 知能 入門
Preferences
Google
100
Yahoo!
0
Search Results for 人工 知能 入門 Google: about 39,100 matches Yahoo: about 0 matches
Semi-Structured Scrapbook
B89 · 2005/05/16 02:41 [edit]. metadata. supervisor: 長尾眞; publisher: 近代科学社; title:
fhuman.esys.tsukuba.ac.jp/~nozom/s3/?id=11158028290178 - 3k - Cached
Score = 100.00 Google rank = 1
人工知能入門
Information & Computing 107
www.saiensu.co.jp/books-htm/ISBN4-7819-1136-6.htm - 6k - Cached
Score = 50.00 Google rank = 2
Amazon.co.jp:考えるコンピューター—人工知能入門: 本
考えるコンピューター
www.amazon.co.jp/exec/obidos/ASIN/4764901080 - 31k - Cached
Score = 33.33 Google rank = 3
岡田斗司夫的「學入門」
... 學
bbs.ee.ntu.edu.tw/boards/Comics/11/36/3.html - 8k - Cached
Score = 25.00 Google rank = 4
續岡田斗司夫的「學入門」
發信人: Alplus.bbs@aidebbs (
bbs.ee.ntu.edu.tw/boards/Comics/11/36/4.html - 4k - Cached
Score = 20.00 Google rank = 5
長尾研教授担当講義
www.nlab.sogo1.ynu.ac.jp/ynu/kougi/kougi.html - 8k - Cached
Score = 16.67 Google rank = 6
Mana House 詳細内容:人工知能入門
ジンコウ チノウ ニュウモン
www.manah.net/book/product.jsp?sku=B4781911366 - 37k - Cached
Ads by Goooooogle
ホイクカフェ
保育関係者専門のコミュニティサイト 仕事のことから趣味のことまで
hoikucafe.jp/
プログラミングの通信講座
プログラムの入門から応用まで学べる 総合ITスクール、札幌情報技術学院
www.sitc.ac
ネット証券を賢く使う方法
無料情報の活用で時間とお金を節約。投資スタイル、目的別で会社を選ぶ。
www.best-investor.com/
大学受験に役立つ情報満載
国公立・私立大学情報が充実! ネットで過去問や大学資料も入手可能
www.52school.com/
日経BPが徹底解説。
ファイナルファンタジーで プログラミングを始めよう。
itpro.nikkeibp.co.jp/a/it/pl
Internet
100%


人工 知能 入門 - Releton Search - Windows Internet Explorer
http://www.releton.com/?q=%E4%BA%BA%E5%B7%A5%E3%80%80%E7%9F%A5%E8%83%BD%E3%80%80%E5%85%A5%E9%96%80
File Edit View Favorites Tools Help
Google 検索 PageRank 17 をブロックしました チェック オプション
Yahoo! JAPAN
人工 知能 入門 - Releton...
Page Tools
Web Images Maps Keywords Text
Releton
Web
Search
人工 知能 入門
Preferences
Google
Yahoo!
0
100
Search Results for 人工 知能 入門 Google: about 39,100 matches Yahoo: about 0 matches
http://www.jcedu.org/fxzd/jgj/txt/jz/qt/jingang39.txt
金剛般若波羅蜜經講義 震旦淸信士勝觀江妙熙 遺著 附:金剛經校勘記 (本講義以再版補足之本,與江居士親筆原稿,覆校重刊。)
www.jcedu.org/fxzd/jgj/txt/jz/qt/jingang39.txt
Score = 100.00 Yahoo rank = 1
Nintendo World BBS GBA|NDS|PSP - Robot Town - あの...名付けて、『ブルートウ戦紀』![完結]
Nintendo World BBS gba,emu,wsc,psp,nds,nintendo,sony,rom,任天堂,索尼,模拟器,下载,攻略,
笨机 - Discuz! Board ... 第一章 入門 冥王之初擁. 既然是入門,那麼這部分只是針對SRW初心者的,各位SRW老飯絲可以先 ...
bbs.newwise.com/viewthread.php?tid=95344 - 240k - Cached
Score = 50.00 Yahoo rank = 2
Magic-Pro (Singapore) - Awards
... 200G核心更能發揮,效能完全抛離其他K8 IGP平台,而價錢亦十分合理,絕對是3D入門級K8最佳平台 ...
magic-pro.com.sg/content/view/33/49
Score = 33.33 Yahoo rank = 3
圣水寺圣域佛教网大藏经文本
圣水寺圣域佛教网大藏经文本
a.heshang.net/dzzhtml/T46-Fh/1912.txt.html
Score = 25.00 Yahoo rank = 4
學佛入門- Powered by 博易-AnyP.cn
欢迎光临..., 博易博客,简单,时尚,实用,好玩,还免费.每个人都应该有一个博易博客
lbpet.anyp.cn/rumen/articles/040925100012828.aspx?z=143929&m=271833
Score = 20.00 Yahoo rank = 5
佛學問答類編通問第一『雪廬老人專輯』念佛网 - powered by phpwind.net
... 兩類皆不同,即上兩類聖,所知所能者,佛皆知能,佛所知所能者,上兩類聖,有所不知不能 ...
www.nianfo.net/bbs/simple/index.php?t794.html
Score = 16.67 Yahoo rank = 6
Ads by Goooooogle
ホイクカフェ
保育関係者専門のコミュニティサイト 仕事のことから趣味のことまで
hoikucafe.jp/
プログラミングの通信講座
プログラムの入門から応用まで学べる 総合ITスクール、札幌情報技術学院
www.sitc.ac
ネット証券を賢く使う方法
無料情報の活用で時間とお金を節約。投資スタイル、目的別で会社を選ぶ。
www.best-investor.com/
大学受験に役立つ情報満載
国公立・私立大学情報が充実! ネットで過去問や大学資料も入手可能
www.52school.com/
日経BPが徹底解説。
ファイナルファンタジーで プログラミングを始めよう。
itpro.nikkeibp.co.jp/a/it/pl
Internet
100%


人工 知能 入門 - Releton Search - Windows Internet Explorer
http://www.releton.com/?q=%E4%BA%BA%E5%B7%A5%E3%80%80%E7%9F%A5%E8%83%BD%E3%80%80%E5%85%A5%E9%96%80
File Edit View Favorites Tools Help
Google 検索 PageRank 17 をブロックしました チェック オプション
Yahoo! JAPAN
人工 知能 入門 - Releton...
Page Tools
Web Images Maps Keywords Text
Releton
Web
Search
人工 知能 入門
Preferences
Google
50
Yahoo!
50
Search Results for 人工 知能 入門 Google: about 39,100 matches Yahoo: about 0 matches
Semi-Structured Scrapbook
B89 · 2005/05/16 02:41 [edit]. metadata. supervisor: 長尾眞; publisher: 近代科学社; title:
fhuman.esys.tsukuba.ac.jp/~nozom/s3/?id=11158028290178 - 3k - Cached
Score = 100.00 Google rank = 1
人工知能入門
Information & Computing 107
www.saiensu.co.jp/books-htm/ISBN4-7819-1136-6.htm - 6k - Cached
Score = 83.33 Google rank = 2
http://www.jcedu.org/fxzd/jgj/txt/jz/qt/jingang39.txt
金剛般若波羅蜜經講義 震旦淸信士勝觀江妙照 遺著 附:金剛經校勘記 (本講義以再版補足之本,
與江居士親筆原稿,覆校重刊。)
www.jcedu.org/fxzd/jgj/txt/jz/qt/jingang39.txt
Score = 83.33 Yahoo rank = 1
Amazon.co.jp:考えるコンピューター人工知能入門: 本
考えるコンピューター
www.amazon.co.jp/exec/obidos/ASIN/4764901080 - 31k - Cached
Score = 71.43 Google rank = 3
Nintendo World BBS GBA|NDS|PSP - Robot Town - あの...名付けて、『ブルートウ戦紀』![完結]
Nintendo World BBS gba,emu,wsc,psp,nds,nintendo,sony,rom,任天堂,索尼,模拟器,下载,攻略,
掌机 - Discuz! Board ... 第一章 入門:冥王之初擁. 既然是入門,那麼這部分只是針對SRW初心者
的,各位SRW老飯絲可以先 ...
bbs.newwise.com/viewthread.php?tid=95344 - 240k - Cached
Score = 71.43 Yahoo rank = 2
岡田斗司夫的「學入門」
... 學
bbs.ee.ntu.edu.tw/boards/Comics/11/36/3.html - 8k - Cached
Score = 62.50 Google rank = 4
Ads by Goooooogle
ホイクカフェ
保育関係者専門のコミュニティサイト 仕事のことから趣味のことまで
hoikucafe.jp/
プログラミングの通信講座
プログラムの入門から応用まで学べる 総合ITスクール、札幌情報技術学院
www.sitc.ac
ネット証券を賢く使う方法
無料情報の活用で時間とお金を節約。投資スタイル、目的別で会社を選ぶ。
www.best-investor.com/
大学受験に役立つ情報満載
国公立・私立大学情報が充実! ネットで過去問や大学資料も入手可能
www.52school.com/
日経BPが徹底解説。
ファイナルファンタジーで プログラミングを始めよう。
itpro.nikkeibp.co.jp/a/it/pl
Internet
100%

はじめに(2/2)

- Metalearning
  - IRアルゴリズムの組合せ方法を学習
  - 機械学習による組合せモデルの構築
- 新規性
  - 従来研究では、固定の組合せ関数を利用
  - 個別データに対して複数のIRアルゴリズム
  - 非線形な組合せ関数も利用可能

従来研究(1/2)

- スコアの組合せ関数の利用 [Lee 95]
  - 2～6個のアルゴリズムの組合せ
  - 1つの質問に対し、各アルゴリズムを文書集合に適用
  - 各アルゴリズムのランキングを作成
  - 組合せ関数でひとつのランキングにまとめる
- 結合式
  - 個別スコア、スコアの数、アルゴリズム数を使った関数

従来研究(2/2)

- 複数の重み付け手法の組合せ [Lee 95,97]
  - 単語とフレーズの組合せ
- 学習アルゴリズムの組合せ [Hull 96]
  - Rocchioの検索質問拡張に適用
  - 最近傍法、線形判別、ニューラルネットワーク
  - 組合せにより性能が改善
- 最適な検索手法を選択するメタモデル [Mayfield 00]
  - 文書に依存した単語、フレーズの獲得を目指す

ベクトル空間モデル

- 文書や検索質問をベクトル空間内の「点」と表現
  - t_i は重み付け手法により数値化

$$\mathbf{d_a} = (t_1, t_2, \cdots, t_n)$$

- d_a, d_b 間の類似度を計算

$$\mathrm{similarity}(\mathbf{d_a}, \mathbf{d_b})$$

- 類似度の大きい順に並び替え、ランキング作成

重み付け手法

term frequency	tf
word count	tf/wc
log	log(tf + 1)
binary	0 or 1
maximum normalization	tf / TF
average normalization	tf / (TF – ave(tf))
TF*IDF	tf * (N / n)
TF*ln(IDF)	tf * ln(N / n)

類似度計算手法

- ユークリッド距離の逆数 $1/\left(\sqrt{\sum\left((v1_i - v2_i)^2\right)}+1\right)$
- マンハッタン距離の逆数 $1/\left(\sum\left(v1_i - v2_i\right)+1\right)$
- 要素最大値の逆数 $1/\left(\max\left(v1_i - v2_i\right)+1\right)$
- ダイス(Dice)係数 $2\times w/\left(n1+n2\right)$
- ジャカード(Jaccard)係数 $w/\left(N-z\right)$

Metalearning

- 機械学習
  - 学習データ: (x, y) からなるデータ集合
    - x : n 次元ベクトル
    - y : 真理値 $y \in \{0, 1\}$
  - 学習データを満足するモデル(関数) f を見つける

$$y = f(x) = f_a(x) = f(x; a)$$

    - a : パラメータ(正規分布ならば平均と分散 など…)
  - f の用途: 未知データ v の真理値を求める
    - y = f(v)

Metalearning

- モデル作成法(1) [Grossman 96]
  - 複数の学習データ $\{L_1, \cdots, L_n\}$ からモデル構築
  - n 個のモデル $\{f_1, \cdots, f_n\}$ が得られる
  - モデルの選択は多数決

Metalearning

- モデル作成法(2) [提案法]
  - 学習データ L を n 個複製 $\{L_1, \cdots, L_n\}$
  - 各学習データから異なるモデルを構築 $\{f_1, \cdots, f_n\}$
  - テストデータ v から n 個のスコアを計算
    $\{f_1(v), \cdots, f_n(v)\}$
  - 各モデルを結合する結合モデルを構築
    - 多項式モデル
    - 回帰モデル など…

Implementation

- PATTERN システムへ Metalearning を組込む
  - アンサンブル学習を主としたデータマイニングソフト
- 文書集合からモデルの学習
  - 文書集合から文書-索引語行列を作成
  - 索引語 : 単語、または *n*-gram
  - stop words, stemming の処理も行う

実験データ

- TREC3 FBISデータ
  - FBIS (Federal Broadcast Information Service)
  - 300件の文書
  - 29件の検索質問
  - 8700(＝300×29) 通りの文書-質問の組合せ
  - テストデータより、組合せのうち 110 件が正解

スコアの計算方法

- 各文書-質問のペアについて40通りの類似度計算
  - 5種類の類似度関数
  - 8種類の重み付け手法
- 類似度の結合手法
  - 線形回帰モデル
  - 2種類の2次回帰モデル
  - 3次回帰モデル
- ベースライン
  - 単独IRアルゴリズムで最も良い性能をもつ手法の精度

実験結果

No. Rel.	Baseline	Linear	Cubic	Quad1	Quad2
10	592	380	145	75	76
20	1352	931	299	212	195
30	1825	1417	820	400	372
40	2453	1820	1122	670	622
50	4013	2306	1422	933	920
55	4383	2442	1598	1168	1137
60	5044	2601	1730	1458	1345
70	5817	3136	2101	1910	1734
80	6733	4164	2775	2668	2312
90	7420	4880	3472	3645	3035
100	8328	5552	5211	4875	5353
110	8644	7597	8101	7780	7948

考察

- 実験結果の表より、
  - 精度がベースラインを大幅に向上
- 再現率 0.09 (= 10/110) での精度
  - Baseline：0.016 (= 10/592)
  - 2次回帰1：0.133 (= 10/75)　(700%増！)
- 弱い手法の組合せで強い手法を作ることが可能

追加実験

- 文書数の拡張
  - 11,188件の文書
  - 10件の検索質問
- 54通りのIRアルゴリズム

- IRアルゴリズム数増加は無意味
  - IRアルゴリズム同士が類似
  - 28通りのアルゴリズムしか使われない
  - 基本的なアルゴリズムだけ残る

追加実験

- 組合せ関数の置き換え
  - 決定木
  - ロジスティック回帰
  - Random Forest(木構造に対するアンサンブル手法)
- 予備実験の結果
  - どの機械学習手法も単独利用より性能が良かった
  - 基本IRアルゴリズムに高い重み
    - 一般的に有名な重み付け手法、類似度計算手法とは限らない

おわりに

- 検索アルゴリズムを組合わせたMetalearning
  - 組合せにより、精度が大幅に改善
- 学習データからモデルを構築
  - 複数のIRアルゴリズムを組合わせる
  - 単純な非線形の組合せモデルが最も有効
  - 機械学習など、他の手法への可能性
- 小規模なデータでの学習結果
  - 他の大規模なデータに対応できないのか？