

2009年9月4日（金）

筑波大学知的コミュニティ基盤研究センター 研究談話会

高信頼性情報の取得を目指した ブロガー熟知度に基づく ブログランキングシステム

中島 伸介
(京都産業大学コンピュータ理工学部)

背景

- ブログの検索やランキングに対する要求の増大
- Web検索のランキングアルゴリズムをそのままブログ検索に適用することは困難

一般的なWebページ：

価値の高いページには数多くのページからリンクが張られている。

ブログエントリー：

価値の高いブログエントリーだとしても
投稿直後には外部からのリンクは存在しない

ブログの特長である即時性を失うことのない
ランキング方式が必要

従来手法

- ブログエントリーに対するランキング

- 新着順や、キーワードとの関連性に基づくランキングが多い。
(Googleブログ検索等)

コンテンツの価値を
評価できない

- ブログサイトに対するランキング

- リンク数, アクセス数, 投票等によるランキングが多い。
(Technorati等)

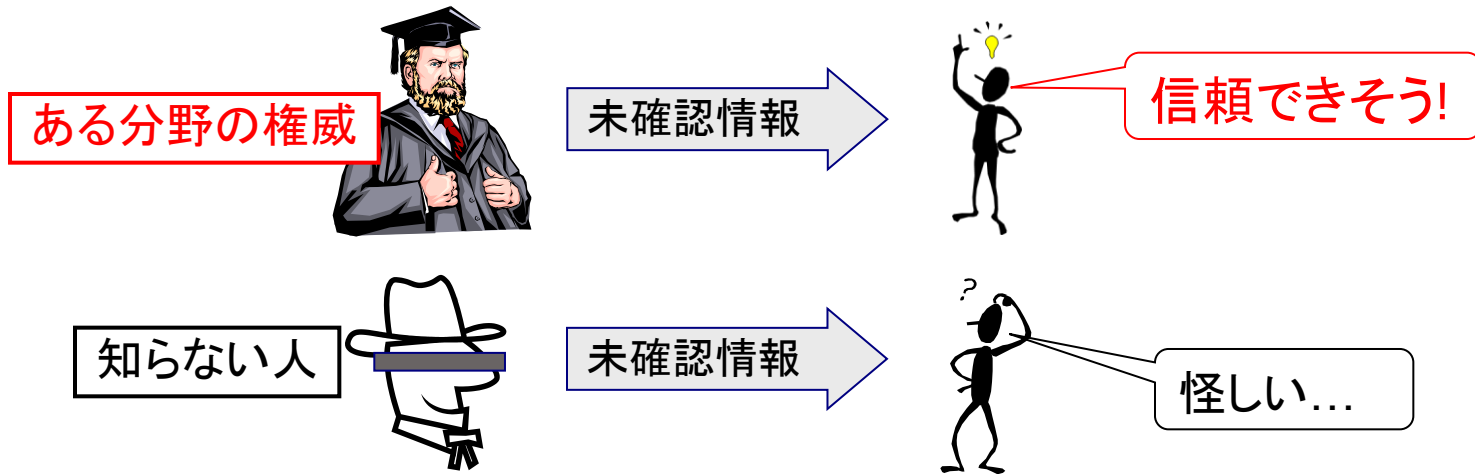
目的のトピックに関する
最新エントリーが存在する
とは限らない

(無名のサイトは上位にランキングされにくい)

体験熟知度に基づくブログランキング

- 基本的な考え方

- 特定トピックに関して、熟知度が高いブロガーが書いた記事は、素人が書いた記事よりも価値が高い。



体験熟知度に基づくブログランキングとは...

“対象トピックに関して詳しく書かれたエントリ”を数多く投稿したブロガーに対しては、このトピックに関する熟知度を高いと判定し、この熟知度を用いてブログエントリのランキングを行ったもの。

本研究の目的

- ブロガーが過去に投稿したエントリにて使用している“ある話題を表すキーワード群”の頻度、網羅率、専門性から、その話題に対するブロガーの熟知度を算出し、これに基づいてブログエントリのランキングを算出しようとするブログランキングシステムを開発する。

実証実験システム (http://kizasi.jp/labo/nict_h20/)

● kizasi・京都産業大学 共同研究プロジェクト

本機はNICT委託研究にて一般ブロガーと熟知ブロガーを比較することで、信頼性の高い情報により容易に、より速くたどり着くことを目的とするシステムです。
実験・検証を行った後、新しいサービスとして追加するかどうかを検討します。 [詳細アンケートを募集しています。](#)

キーワードを入力

検索

おすすめキーワード

松野佳 織田信長 葉 南雄 フレグランス 宇宙 バチンコ

● ダイエット 熟知グループ解析

関連熟知グループ

ダイエット

健康

サプリメント

食事

フィットネス

美容

仕事

熟知ブロガーランキン

全ブロガー 熟知ブロガー

● 熟し日記 「抜け毛、髪毛防止対策に効果的!?」誰でもできるツヤのある髪になる方法

2009年3月17日 20:18

ダイエット(37.2%) エクササイズ(23.2%) 美容(18.1%) 食事(11.3%) フィットネス(9.9%)

シャンプー自体の成分にも、頭皮に重要な影響を与えるものもあります。これらを予防するためには、**健康**な日常生活を心がけた上で、マッサージなどの対策を行うことが大切です。しかし、一層の対策と予防は、心身ともに健康な生活を送ることだということを忘れないでください。

[readinggood.blog.shinichi.jp/Entry/552/](#) ● このブログを読む

● 3月19日 僕の髪にも効果

2009年3月10日 09:58

ダイエット(34.6%) フィットネス(27.4%) ウォーキング(17.9%) コーヒー(10%) 美容(9.9%)

私は完璧な「下手の横好き」ですが、ゲーマーでした。今は、高いし、あんまり**健康**じゃない(笑)ので我慢してたのですが、思い切って買っちゃおうかな。

[3kuyokanomama.blog.shinichi.jp/Entry/552/](#) ● このブログを読む

NEXT

● 全ブロガー解析

全成分

感情 不快・不愉快 役所

キーワード

心 話 多い 感じ 行く よく 日本 風 邪 悪い 情報 食べる 人間 家 本当に 高い ブログ 女性 方法 お金 無い

● 悪い感じ ● 宿しや寝い ● 不快・不愉快 ● 落ち込み ● イライラ ● 呆れる ● 嫌しい ● 差別 ● 悲しい ● 恐怖 ● 怒り

● 面白い ● 感謝 ● 面白い ● スッキリ ● おもしろい ● 尊敬

● 宿しや寝い ● 不快・不愉快 ● 落ち込み ● イライラ ● 呆れる ● 嫌しい ● 差別 ● 悲しい ● 恐怖 ● 怒り

全成分

役所 感情 スッキリ

キーワード

所 理由 母 早く 状況 出る 手 行った 結果 気分 外 影響 飲み タイトル コレ ほしい ください いけない 1日 1

熟知グループおよび 熟知度ブログランキング

熟知度ブログランキング

1. 熟知領域リストの作成
2. ブロッガーの熟知度スコアの算出
3. ブログエントリのランキングの算出

熟知領域リスト(および関連語辞書)の作成

- 方法

- 「ファン」「マニア」「フリーク」等のキーワードでWeb検索を行い, これら検索語の直前の語句のうち, 出現頻度が高いものを辞書に登録.

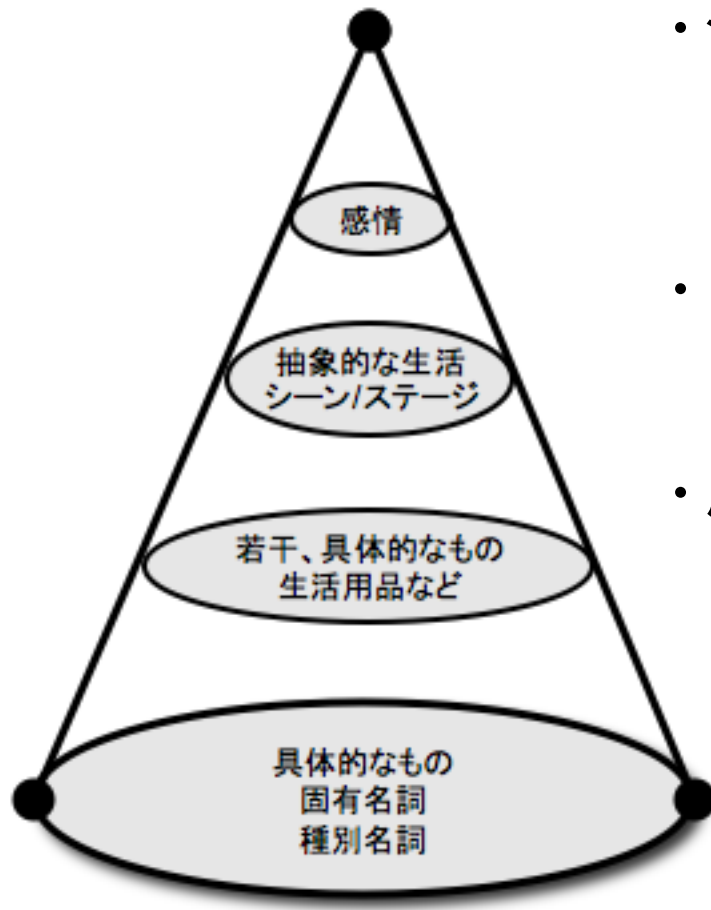
→ 500領域



- 独自に開発した生活体験シソーラス (LETS) を用いて, そのカテゴリを熟知領域リストとして採用.

→ 14000領域

生活体験シソーラス (LETS)



生活体験シソーラスLETSの概念図

- シソーラス

ブログやニュースなどの実テキストにしばしば表現される生活体験を体系的に整理・分類したシソーラス

- 自動的カテゴリ辞書管理システム

ブログデータを用いた連想辞書の自動生成を行うシステム

- 成分解析エンジン

任意の入力テキストを分類するグラデーション・エンジン。広告や商品のレコメンデーションに応用している

ブロガーの熟知度スコアの算出(1/2)

- 対象熟知領域 g に対する、あるブログエントリの関連度スコア $score_g(e)$

$$score_g(e) = \sum_{j=1}^n (\alpha_j \cdot \beta_j \cdot \gamma_j)$$

α_j : 順位 j 番目の共起語の重み β_j : j 番目の共起語の共起度 n : 共起語数 ($n=400$)

γ_j : 共起度順位 j の共起語が当該エントリ内に 存在する場合 1, しない場合 0

ブロガーの熟知度スコアの算出(2/2)

- 対象熟知領域 g に対するブロガー b の熟知度スコア $score_g(b)$

$$score_g(b) = \frac{l}{n} \cdot \frac{\log(m)}{m} \cdot \sum_{i=1}^m score_g(e_i)$$

e_i : ブロガー b が投稿した, エントリ

m : ブロガー b が対象期間内に投稿したエントリ数

l : ブロガー b が対象期間内に投稿したエントリに出現した共起語数

l/n : ブロガー b が使用した共起語の全共起語に対する網羅率

$score_g(e_i)$: エントリ e_i の対象熟知領域 g に対する関連度スコア

別途, スпамフィルタを採用

スパムフィルタの例（１）

このブロガーの総エントリ数:2932

文章の形態素総数:109418

文章を書いた期間:2008/10/01 01:15:00～2009/04/02 15:00:00

時	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Sun	10	15	16	27	19	28	24	20	21	18	17	16	15	19	15	14	20	14	14	11	15	7	6	10
Mon	5	19	18	24	21	29	22	26	21	19	21	21	15	18	20	16	19	20	14	17	12	9	13	10
Tue	10	15	22	25	22	33	23	22	21	21	19	16	14	19	24	16	14	10	12	15	11	11	6	10
Wed	10	16	24	27	23	30	25	25	20	19	20	22	16	23	14	21	18	13	14	12	11	12	11	9
Thu	15	20	19	24	18	31	26	27	22	17	21	20	18	18	19	19	16	18	12	13	14	10	6	9
Fri	10	15	21	22	18	30	23	26	22	18	18	17	11	22	17	13	17	15	15	13	13	7	10	11
Sat	13	16	22	26	19	24	23	23	21	22	19	18	17	17	18	17	19	15	13	16	18	14	10	10

全ての時間帯に異常な数のエントリを投稿

スパムフィルタの例（２）

このブロガーの総エントリー数:402

文章の形態素総数:11685

文章を書いた期間:2008/10/01 07:39:00~2009/03/25 11:08:00

時	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
Sun								14	3			11	1		11	2	1			10				
Mon								21				15	1		12	1		1		10				
Tue								18	3			14	1		9	2				8			1	
Wed						1		22	1			13	1		10	1				10				
Thu				1			1	18	2			9	2		10	2				9	4			
Fri								19	1	1		11	3		9	4				9				
Sat								20	3			10	1		12					10	2			

規則正しく決まった時刻にエントリーを投稿

ブログエントリのランキングの算出

1. 検索キーワードを含むブログエントリを検索すると共に、システムはこれらのエントリを投稿したブロガーを特定し、その熟知度スコアを取得する。
2. これらブロガーが属する熟知領域毎（熟知グループ）に検索したエントリ进行分类する。
3. 各熟知グループに分类されたエントリを、各ブロガーの熟知度スコアに基づいてソートすることで、その熟知グループに関するランキングを行う。
4. 上記により、一度の検索において複数の熟知領域に関するランキングを提供することができる。

熟知グループ選択および検索結果の提示部

- 左カラムでは、検索語と関連性の高い熟知グループをリスト表示。（関連度は、熟知グループの使用語彙により評価。）
- 右カラムでは、全ブロッガーまたは熟知ブロッガーの検索結果を表示。
- スニペット部にブロッガーが属する熟知グループの割合情報を表示。

検索語:「健康」 熟知グループ:ダイエット

○ ダイエット 熟知グループ解析

関連熟知グループ

ダイエット

健康

サプリメント

食事

フィットネス

美容

仕事

熟知ブロッガーランキング

全ブロッガー 熟知ブロッガー

○ 検索日記「抜け毛、薄毛防止対策に効果的！？髪でもできるツヤのある髪になる方法」

2009年3月17日 20:18:04

ダイエット(37.2%) エクササイズ(23.2%) 美容(18.1%) 食事(11.3%) フィットネス(9.9%)

シャンプー自体の成分にも、頭皮に重要な影響を与えるものもあります。これらを予防するためには、**健康的**な日常生活を心がけた上で、マッサージなどの対策を行なうことが大切です。しかし、一番の対策と予防は、心身ともに健康な生活過ごすことだということを忘れないでください。

Feelinggood blog shinobi.jp/entry/520/ → このブログを読む

○ 3/10 実の娘やると...

2009年3月10日 09:58:05

ダイエット(34.6%) フィットネス(27.4%) ウォーキング(17.9%) コーヒー(10%) 菓子(9.8%)

私は完璧な「下手の横好き」ですが、ゲーマーでした。今は、高いし、あんまり**健康的**じゃない(笑)ので我慢してたのですが。思い切って買っちゃおうかなあ。

akuyokanumore blog shinobi.jp/entry/322/ → このブログを読む

NEXT

実証実験システム (http://kizasi.jp/labo/nict_h20/)

kizasi・京都産業大学 共同研究プロジェクト

本機関はNICT委託研究にて一般ブロガーと熟知ブロガーを比較することで、信頼性の高い情報により容易に、より速くたどり着くことを目的とするシステムです。
実験・検証を行った後、新しいサービスとして追加するかどうかを検討します。 [詳細アンケートを募集しています。](#)

キーワードを入力
 検索

🔍 健康

おすすめキーワード
格闘技 織田信長 猿 悪徳 フレグランス 宇宙 パチンコ

ダイエット 熟知グループ解析

関連熟知グループ

- ダイエット
- 健康
- サプリメント
- 食事
- フィットネス
- 美容
- 仕事

全ブロガー
熟知ブロガー

🔍 最も初期 「抜け毛、髪も防止対策に効果的！？誰でもできるツヤのある髪になる方法」

2009年3月17日 20:18

ダイエット(37.2%)	エクササイズ(23.2%)	美容(18.1%)	食事(11.3%)	フィットネス(9.9%)
--------------	---------------	-----------	-----------	--------------

シャンプー自体の成分にも、頭皮に重要な影響を与えるものもあります。
これらを予防するためには、**健康的な日常生活**を心がけた上で、マッサージなどの対策を行なうことが大切です。
しかし、一冊の対策と予防は、心身ともに健康な生活を送ることだということを忘れないでください。

[readinggoal.blogspot.jp/entry/5037/](#) • このブログを読む

🔍 ムネ子 髪の悩める女こそ。

2009年3月10日 09:58

ダイエット(34.6%)	フィットネス(27.4%)	ウォーキング(17.9%)	コーヒー(10%)	子育て(9.8%)
--------------	---------------	---------------	-----------	-----------

私は完璧な「下手の横好き」ですが、ゲーマーでした。
今は、高いい。あんまり**健康的**じゃない（笑）ので我慢してたのですが、思い切って買っちゃおうかな。

[okuyukikonomama.blogspot.jp/entry/5037/](#) • このブログを読む

全ブロガー解析

主成分

脂肪 不快・不愉快 疲労

- 痛む感じ
- 感動
- ラブラブ
- 可愛い
- 美しい
- 感謝
- 面白い
- スッキリ
- ねええい
- 尊敬

主成分

疲労 恐怖 スッキリ

キーワード

心 話 多い 感じ 行く よく 日本 風 邪 悪い 情報 食べる 人間 家 本当に 高い ブログ 女性 方法 お金 無い

- 宿い・寝い
- 不快・不愉快
- 落ち込み
- イライラ
- 呆れる
- 悔しい
- 差別
- 悲しい
- 恐怖
- 怒り

キーワード

所 理 由 母 早く 状況 出る 手 行った 結果 気分 外 影響 飲み タイトル コレ ほしい ください いけない 1日 1

ニュアンス比較部

ニュアンス比較機能の開発

- ・ブロガー全体と、対象熟知ブロガーグループのニュアンスを比較することで、熟知ブロガーグループの立場を理解しやすくする。

全ブロガー

ニュアンスの例

熟知ブロガー

全ブロガー解析

主成分



恐怖



不快・不愉快



疲労

キーワード

心 話 多い 感じ 行く よく 日本 風
邪 悪い 情報 食べる 人間 家 本当に
高い ブログ 女性 方法 お金 無い

- ・ 清い感じ
- ・ 感動
- ・ ラブラブ
- ・ 可愛い
- ・ 美しい
- ・ 感謝
- ・ 面白い
- ・ スッキリ
- ・ おいしい
- ・ 尊敬
- ・ 怪しい・疑い
- ・ 不快・不愉快
- ・ 落ち込み
- ・ イライラ
- ・ 呆れる
- ・ 悔しい
- ・ 疲労
- ・ 悲しい
- ・ 恐怖
- ・ 怒り

主成分



疲労



恐怖



スッキリ

キーワード

所 理由 母 早く 状況 出る 手 行った
結果 気分 外 影響 飲み タイトル
コレ ほしい ください いけない 1日 1

ニュアンス比較機能の開発

＜判定方法＞

- ・各ニュアンスに対して400の関連語をリストアップし、この関連語の使用状況に基づいて各ブログエントリのニュアンスを判定.

全ブロッガー

全ブロッガー解析

主成分



恐怖



不快・不愉快



疲労

キーワード

心 話 多い 感じ 行く よく 日本 風
邪 悪い 情報 食べる 人間 家 本当に
高い ブログ 女性 方法 お金 無い

ニュアンスの例

- ・ 清い感じ
- ・ 感動
- ・ ラブラブ
- ・ 可愛い
- ・ 美しい
- ・ 感謝
- ・ 面白い
- ・ スッキリ
- ・ おいしい
- ・ 尊敬
- ・ 怪しい・疑い
- ・ 不快・不愉快
- ・ 落ち込み
- ・ イライラ
- ・ 呆れる
- ・ 悔しい
- ・ 疲労
- ・ 悲しい
- ・ 恐怖
- ・ 怒り

熟知ブロッガー

主成分



疲労



恐怖



スッキリ

キーワード

所 理由 母 早く 状況 出る 手 行った
結果 気分 外 影響 飲み タイトル
コレ ほしい ください いけない 1日 1

実証実験システムに関するデータ

- 登録ブロガー数 : 7, 4 2 2, 0 0 0
 - 登録エントリ数 : 1 7 4, 0 0 0, 0 0 0
 - 熟知グループ数 : 1 4, 0 0 0
 - 熟知ブロガー数 : 1 0 0, 0 0 0
- (2009年4月20日現在)

評価実験

評価実験項目

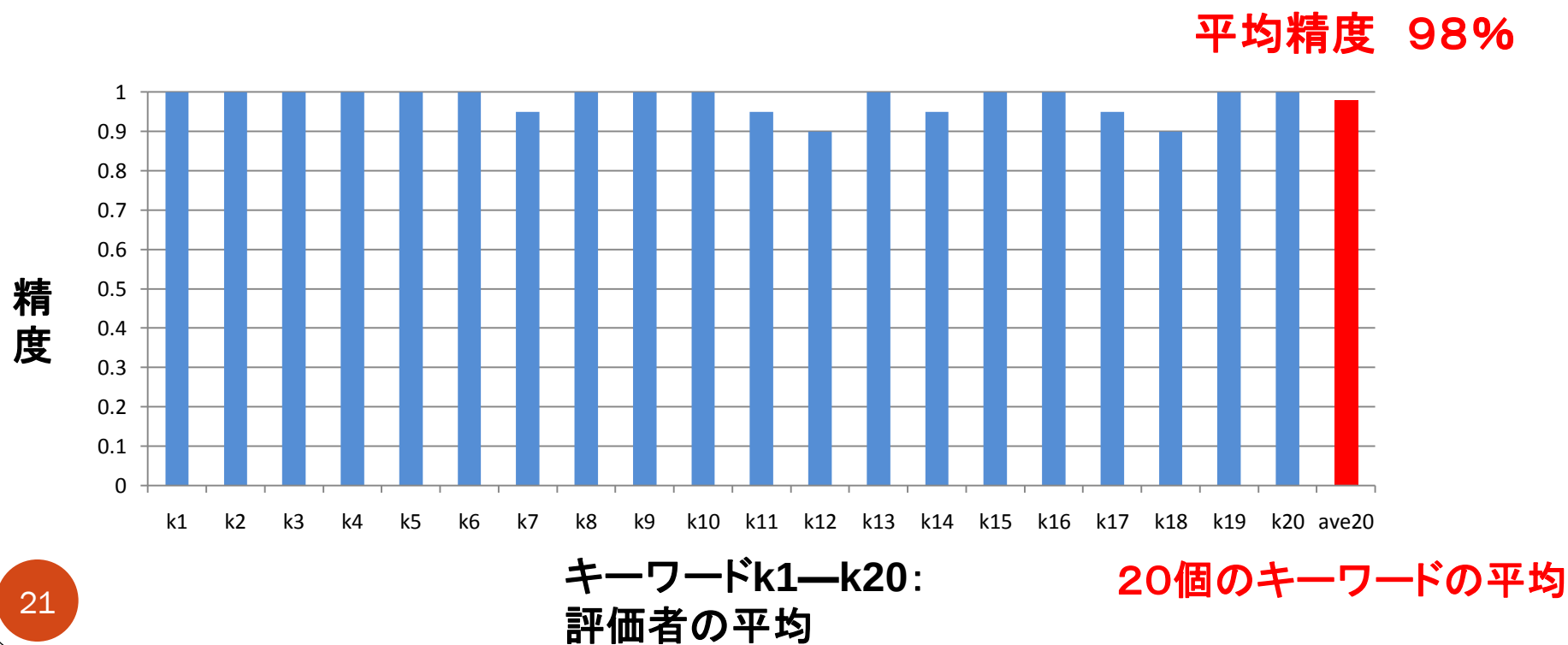
1. 熟知グループの妥当性評価
2. 熟知ブロガーの妥当性評価
3. 熟知度に基づくランキング結果の信憑性評価

使用した検索キーワード

1	Jリーグ	2	GUNDAM	3	スイーツ	4	格闘技	5	鉄道
6	ジャニーズ	7	株	8	酒	9	サッカー	10	野球
11	競馬	12	美術	13	サザンオール スターズ	14	劇団四季	15	アイドル
16	声優	17	iPod	18	ソムリエ	19	Linux	20	健康

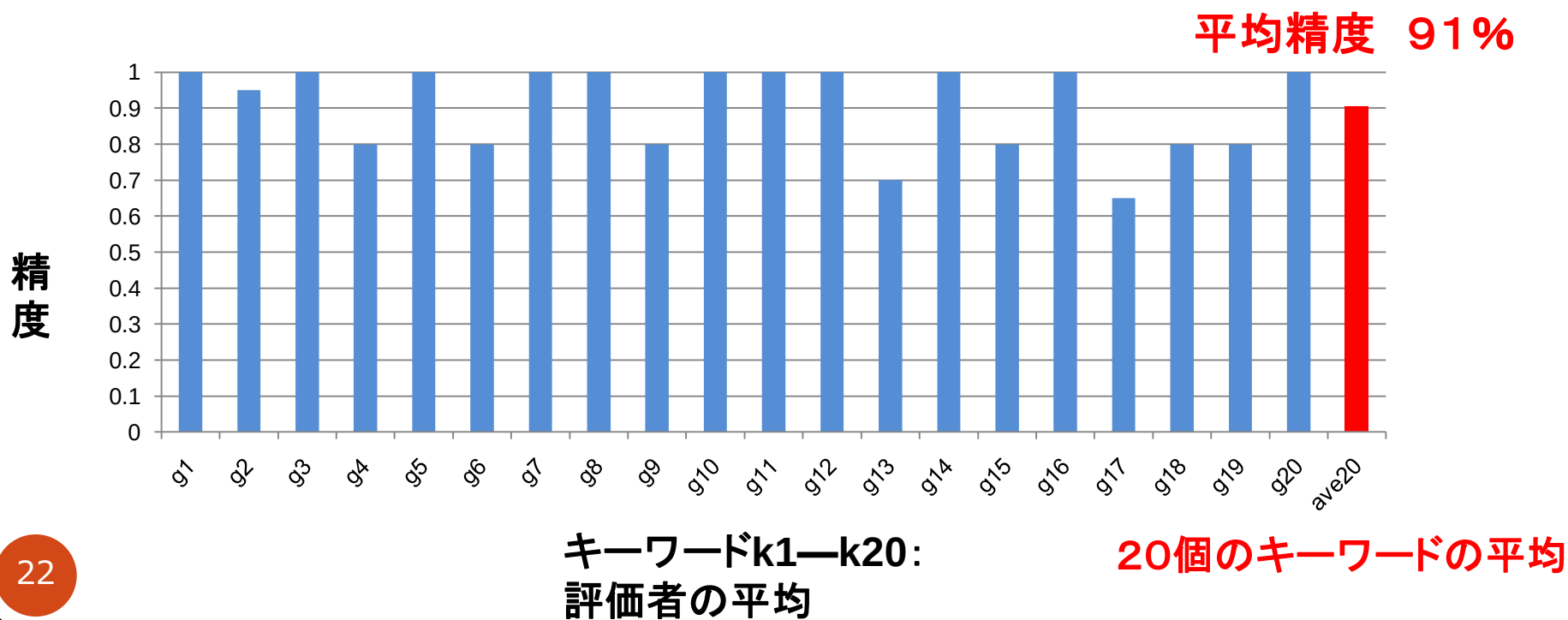
1. 熟知ブロガーグループの妥当性評価

- 検索キーワードに対して表示される熟知ブロガーグループ（上位5個）が妥当かどうかを評価.



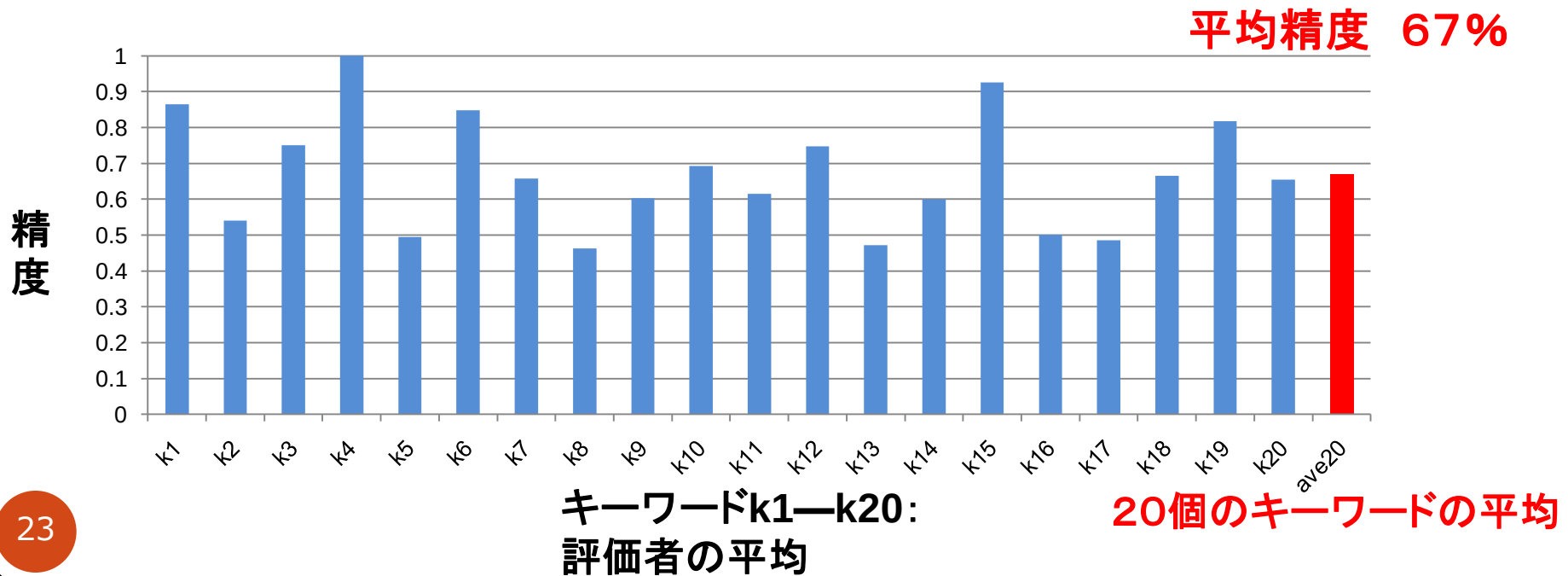
2. 熟知ブロガーの妥当性評価

- 検索キーワードに対する1位の熟知ブロガーグループのランキングを対象とし、上位5人のブロガーが熟知ブロガーとして適当であるかどうかを判定.



3. 熟知度に基づくランキング結果の信憑性評価

- 熟知度ランキング上位5件のブログエントリが信頼できるものであるかどうかを判定.



まとめ

- 過去に投稿したエントリを解析することで得られる各ブロガーの熟知度に基づいて算出されるブログランキング方式を実装し公開した。
また、本システムを用いて評価実験を行い、提案手法の有効性を確認した。
- 今後は、本手法を利用した信頼性の高い情報提示について研究を進める予定である。