修 士 論 文 概 要 書

2010 年　2 月提出

専攻名 （専門分野）	情報理工学専攻	氏 名	黒木さやか	指 導 教 員	山名早人教授　　印
研究指導名	並列・分散アーキテクチャ	学籍番号	CD 5108B040-7		
研 究 題 目	アンカーテキストとリンク構造を用いた同義語抽出手法				

1. はじめに

Web2.0 に代表される新しい情報発信の仕組みにより，企業や商品に対する一般ユーザの評価は，他の一般ユーザだけではなく，企業にとっても貴重な情報源となっている．しかし，企業や商品の評価に関する Web ページは，それらの略称や俗称を用いて書かれていることが多く，検索クエリに正式名称を入力しただけでは取得することができない．そこで本論文ではアンカーテキストとリンク構造を用いることで，略称や俗称などにも対応した同義語抽出の手法を提案する．関連研究としてクエリの翻訳語を発見する研究が存在するが，同手法により作成される翻訳語ランキングは翻訳語をトップにすることを目的としており，頻出語が上位にランキングされるようになっている．従って，頻出ではない略称や俗称などの同義語を効率的に抽出することは難しい．提案手法では，既存手法よりも多くの同義語を抽出すると同時に，新しい同義語候補ランキングの指標を提案し，同義語抽出の効率化を試みる．

2. 関連研究

2.1. クエリ拡張

大量のデータから，検索クエリに関連する文書を探す時，検索クエリと同様の概念を持つ語についても，文字列検索を行うことが効果的であると考えられる．1990 年代までのクエリ拡張分野では，自然言語処理に基づく研究が一般的であった．近年ではインターネットの普及により，シソーラスを利用したクエリ拡張技術[1]や，検索エンジンのクエリログを利用したクエリ拡張技術[2]が研究されている．

2.2. コミュニティ抽出

Web から特定の事柄に関するページ群を取り出す手法として，コミュニティ抽出の研究が挙げられる．[3][4]の研究では，「同じ事柄を述べたページ群は相互リンクを張りやすい」という考え方に基づき，Web のリンク構造から完全，または密な 2 部グラフを抽出している．また，コミュニティ内のリンク数が，コミュニティ外のリンク数よりも多いという定義に基づき，Web のリンク構造に s-t 最大フロー問題を適用した研究もある[5].

2.3. リンク構造を用いた研究

提案手法と同様に，アンカーテキストとリンク構造を用いた研究として，クエリ翻訳[6]が挙げられる．[6]では，クエリと同じ文字列のアンカーテキストがリンクする URL 群に対し，最もリンクしているアンカーテキストを翻訳語としている．[6]によるアンカーテキストの類似度は，式(1)で表される．

$$P(T_s \leftrightarrow T_t) = \frac{\sum_{i=1}^{n} P(T_s|U_i)P(T_t|U_i)P(U_i)}{\sum_{i=1}^{n}[P(T_s|U_i)+P(T_t|U_i)-P(T_s|U_i)P(T_t|U_i)]P(U_i)} \quad (1)$$

- P(Ts|Ui), P(Tt|Ui):アンカーテキスト Ts, Tt から Ui へのリンク数／URL Ui の in-link 数
- P(Ui): URL Ui の in-link 数 ／Web 上の全リンク数（HITS による値）
- n: Web 上の全 URL 数

図1　同一 URL にリンクするアンカーテキスト

3. 既存研究の問題点と解決策

3.1. 提案手法で抽出する同義語

既存研究の問題点を述べる前に，提案手法により抽出する同義語について述べておく．まず，ユーザが特定の企業や人に関する同義語を抽出する際，この企業や人を「対象物」と呼ぶことにする．提案手法で抽出する同義語とは，この対象物を連想できる全ての語である．以下に例を挙げる．

- 対象物の正式名称，正式な略称，翻訳語
- 対象物の一般的な俗称
- 明らかに対象物であると分かる語

3.2. 既存研究[6]の問題点

既存研究では翻訳語がランキングトップになれば良く，頻出ではない語の類似度は低く計算されてしまう．これは，URL 側から見たリンク確率を類似度計算に用いているため，アンカーテキストが他の URL へリンクしている情報を全く活用できないからだと考えられる．4 節では，アンカーテキスト側から見たリンク確率を用いることで，頻出ではない同義語も上位にランキングすることができる，新しい類似度指標を提案する．

3.3. Web のリンク構造に関する問題点

更に精度と網羅性を向上させるために，Web のリンク構造が持つ問題点と解決策について 3 つに分けて説明する．

- 全ての関連 URL を抽出できていない
  対象物に関連する全ての URL に，クエリと同じ文字列のアンカーテキストがリンクしているとは限らず，抽出できない同義語が存在している．同義語がリンクしている URL にリンクするアンカーテキストは，全て同義語候補とする．
- URL の分散により，類似度が低下する
  トップページが複数存在する場合，どのトップページにリンクするかにより同義語の類似度が変化してしまう．トップページのバリエーションを 1 つの URL にまとめることで，同義語の類似度を上げることが望まれる．
- 誤ったリンク情報により同義語候補が増大する
  対象物とは関係のない URL に，クエリからのリンクが存在する場合がある．これらの URL にリンクするアンカーテキスト群は，全て同義語候補として抽出されてしまい，同義語候補ランキングの項目数を増やすことにつながる．誤ったリンク情報を削除することで，同義語候補数を削減することが望まれる．

4. 提案手法

提案手法では，アンカーテキスト側から見たリンク情報を利用する，新しい類似度指標を用いることで，既存研究[6]の問題について解決する．また，Relevance-Feedback の技術を用いることにより，Web の誤ったリンク情報を補正し，同義語ランキングのリランキングを行う．

4.1. 共起強度

アンカーテキスト側から見たリンク確率を利用することで，頻出ではない同義語も上位にランキングできる，新しい類似度指標を提案する．新しい類似度の指標は共起強度と呼び，以下の式で表される．

共起強度 $co(a,b)=\dfrac{2}{\dfrac{1}{P(b|a)}+\dfrac{1}{P(a|b)}}$ (3)

条件付き確率 $P(y|x)=\dfrac{\sum_{u\in c(x,y)} frq(x|u)}{frq(x)}$ (4)

- ✧ frq(x):アンカーテキスト x の総リンク数
- ✧ frq(x|u):アンカーテキスト x から URLu へのリンク回数
- ✧ c(x, y):アンカーテキスト x と y が共通してリンクする URL 群

アンカーテキスト a と b の共起強度は，a と b それぞれの条件付き確率を調和平均したものである．相加平均ではなく調和平均を用いることで，a と b の条件付き確率に差がある場合，最終的な共起強度の値を低く計算することができる．条件付き確率 P(y|x)は，アンカーテキスト x のリンクについて，x と y が共通してリンクする URL へのリンク確率を示している．共通する URL 数ではなく，URL へのリンク確率を用いて共起強度計算を行うため，クエリと同じ文字列のアンカーテキストから多くリンクされる URL に，重みがついた式になっている．

4.2. Relevance-Feedback 技術を用いたリランキング

3.3 節でまとめたように，精度と網羅性の高い同義語抽出を行うためには，Web のリンク構造に関する問題点を解決する必要がある．提案手法では，Relevance-Feedback の技術を利用することでリンク情報の補正を行い，新しいリンク情報を用いて同義語ランキングをリランキングする．リランキングは以下のプロセスで行う．

①同義語ランキングに対する人手による評価

共起強度による同義語ランキングの Top-n に対し，対象物の同義語と思う場合には○を，異なる語と思う場合には×をつける．どちらか判断できない場合には，○×をつけないことにする．以後のプロセスでは，○をつけた語を「○アンカーテキスト」，×をつけた語を「×アンカーテキスト」と表現する．

②○アンカーテキストのマージ

対象物の同義語と判断されたアンカーテキストについて，リンク情報をマージする．○アンカーテキストのみがリンクしていた URL を，クエリがリンクする URL 群に追加することで，新しい同義語候補を抽出することができる．

③○アンカーテキストによる URL マージ

複数 URL へのリンク分散を解消するため，対象物に関連する URL をマージする．この処理により，クエリがリンクする URL 群の 1 部にしかリンクしていない同義語について，共起強度の値を高く計算することができる．マージする URL は，以下の条件を満たすものである．

- ● ○アンカーテキストからのリンク確率の合計が，一定以上となる URL

④×アンカーテキストによるリンク情報の削除

対象物とは関係のないアンカーテキストを×アンカーテキストとして指定することで，誤ったリンク情報を削除する．クエリから対象物とは関係のない URL へのリンク情報を削除することにより，その URL にリンクするアンカーテキスト群を同義語候補から取り除くことが可能である．リンク情報を削除する URL は，以下の条件を全て満たすものである．

- ● ×アンカーテキストとクエリが共通してリンクする URL
- ● URL 側から見たクエリのリンク確率の合計が，一定以下の URL

①～④までのプロセスを繰り返すことにより，対象物の同義語ランキングの網羅性と精度を上げていく．②アンカーテキストのマージはランキングの網羅性向上に有効であり，③④URL のマージ・削除はランキングの精度向上に有効である．

5. 評価実験

既存研究[6]と，共起強度による同義語ランキング，Relevance-Feedback によるリランキングの比較について，精度を表 1に，再現率を表 2に示す．実験には e-Society のデータを用いている．既存研究に比べ，共起強度を用いたランキングは精度と再現率がともに向上していることが確かめられた．共起強度を 0.1 以上にすれば，Relevance-Feedback を用いたリランキングの場合再現率が 80%程度となり，精度も Top-100 と変わらないことが確認できた．

表1　各手法のランキング精度

	Top-100	Top-200	全て	共起強度 0.1 以上
既存研究[6]	8.1%	5.6%	2.1%	—
共起強度	9.9%	7.2%	2.1%	13.5%
リランキング	11.9%	8.1%	1.4%	12.2%

表2　各手法のランキング再現率

	Top-100	Top-200	全て	共起強度 0.1 以上
既存研究[6]	53.1%	69.0%	95.2%	—
共起強度	63.5%	82.8%	95.2%	69.7%
リランキング	70.7%	87.8%	99.5%	79.8%

6. おわりに

本稿では，対象物の略称や俗称を対象とした同義語抽出の手法について提案を行った．アンカーテキストとリンク構造を用いることで，シソーラスには存在しない同義語を抽出することができた．また既存研究に比べ，アンカーテキストを頻出語に限定しなくても，精度の高い同義語ランキングを作成することに成功した．実験では，精度を保った上で，網羅性を約 15%向上させることができた．今後の課題としては，より精度の高いランキングを行うことである．同義語候補ランキングの中には，同義語に記号がついたアンカーテキスト，または「ホームページ」や「トップページ」などの定型語がついたアンカーテキストが現れている．自然言語処理の技術を取り入れることで，これらの語句を取り除くことが可能であると考える．また，コミュニティ抽出の手法を取り入れることで，誤ったリンク情報の除去を自動化できると考えられる．

参考文献

[1] D.Milne, I.H.Witten and D.M.Nichols: “A Knowledge-Based Search Engine Powered by Wikipedia”, Proc. of the 16th ACM Conf. on CIKM, pp.445-454, 2007.

[2] B.M.Fonseca, P.Golgher and B.Possas: “Concept-Based Interactive Query Expansion”, Proc. of the 14th ACM Conf. on CIKM, pp.696-703, 2005.

[3] S. R. Kumar, P. Raphavan, S. Rajagopalan and A. Tomkins: “Trawling the Web for emerging cyber communities”, J. of Computer Networks, Vol.31, pp.1481-1493, 1999.

[4] P. K. Reddy and M. Kitsuregawa: “An approach to relate the Web communities through bipartite graphs”, Proc. of the 2nd Int'l Conf. on WISE, Vol.1, pp301-310, 2001.

[5] G. Flake, S. Lawrence and C. Giles: “Efficient Identification of Web Communities”, Proc. of the sixth ACM SIGKDD, pp.150-160, 2000.

[6] W.H.Lu, L.F.Chien and H.J.Lee: “Translation of Web Queries Using Anchor Text Mining”, ACM Trans. on Asian Language Information Processing, Vol.1, No. 2, pp.159-172, June 2002.

[7] 文部科学省リーディングプロジェクトe-Society: http://cif.iis.u-tokyo.ac.jp/e-society/