E-014

不要語除去を用いた商品情報の比較支援システム

The comparison supporting system of item information using unnecessary word removal

松井　崇†　　大和田　勇人‡
Takashi Matsui　　Hayato Ohwada

1．はじめに

現在のショッピングサイトではユーザが興味をもった商品を一覧表示するような機能を提供している．この機能の専用ページが用意されており、そのため収集と比較の作業は分断されてしまう．また，そのような Web サイトの多くがそのサイト内の商品同士の比較しか行うことができない．　また、サイトによっては商品タイトルに不要な情報が含まれているために，効率的な比較が行えないという問題点が挙げられる．

本研究では商品タイトルに混在する不要語を除去することにより効率的な比較を促し，複数のショッピングサイトの商品情報の比較を支援するようなシステムを構築することを目的とする．

2．関連研究

島村ら[1]は，単一サイトに対して気になった商品情報を常に表示できるようにすることで収集と比較をシームレスに行えるようなシステムを構築している．吉田ら[2]は，テキスト自動分類システムにおいて，カテゴリー間の出現頻度の違いに注目した手法を述べている．相良ら[3]は，店舗情報検索システムにおいて，不用語辞書を作成し、店舗データベースに登録されている店舗名称に含まれる不要語の効率的除去を行っている．

3．提案システム

提案するシステムは，商品タイトルに含まれる不要語の除去を行うことにより，複数サイトの商品の比較が効率的に行える比較支援システムである．本研究における不要語の定義は、商品タイトルに含まれる商品名、メーカー以外すべてとする。図 1 に例を示した。図 2 に提案システムの構成図を示した．

まず，ユーザにより商品ページから商品情報の抽出を行う．商品タイトルについては不要語辞書を用いた不要語の除去を行ったうえで他の商品情報と共に，データベースに格納する．格納された商品情報はブラウザに用意したサイドバー内に表示される．

商品タイトルから不要語の除去を行い，不要語を除いたシンプルな商品タイトルにすることでユーザにとっての効率的な商品情報の比較を促す．まず不要語辞書を作成し，その辞書を参照することで，商品タイトルから不要な単語の除去を行う．

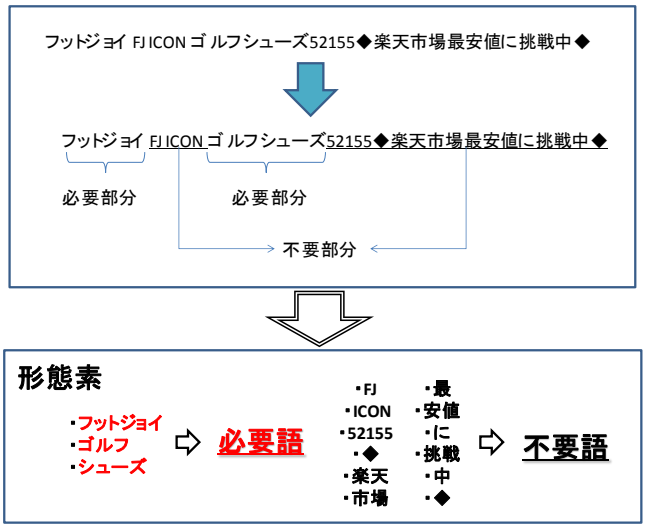

図 1　必要語，不要語

不要語辞書作成は，ショッピングサイトからの商品タイトルの収集，形態素解析，idf 値の算出，閾値の設定による不要語の決定，データベース格納といった流れで行う．idf 値とは，語の特定性を表す尺度の一つであり，式(1)で定義される．

$$\mathrm{idf}_i = \log \frac{|D|}{|\{d : d \ni t_i\}|} \quad (1)$$

$|D|$ は総ドキュメント数，$|\{d : d \ni t_i\}|$ は単語 i を含むドキュメント数である．この idf 値を出すことによって，出現頻度の高い単語は異なるジャンルにおいても出現しており，商品タイトルに含まれる必要語（メーカーや商品名）である可能性は低くなる．不要語辞書作成のために，事前に約 1 万件の商品タイトルを抽出した．閾値を変化させた結果得られた最適な閾値は 4 であった．閾値の設定から得た不要語を不要語辞書に追加し，不要語辞書を作成した．

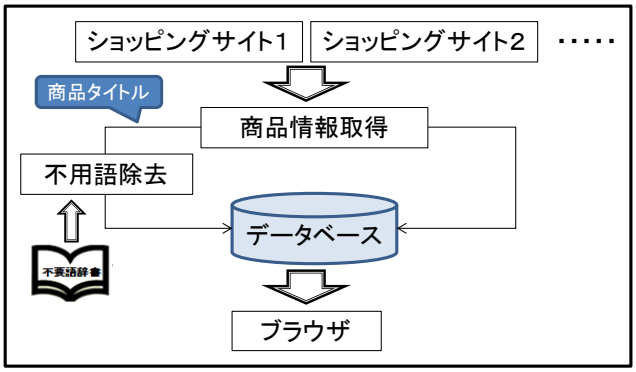

図 2　本システムの構成図

† 東京理科大学理工学研究科経営工学専攻
‡ 東京理科大学理工学部経営工学科

4. 不要語除去性能の評価実験

4.1 実験概要

不要語除去手法の性能評価の実験を行う．評価データとして楽天市場の商品情報を使用した．商品の中から 10 ジャンル全てのジャンルからランダムに選んだ商品 100 品を対象として不要語除去の実験を行う．まず手作業によるラベル付けを行い、各商品タイトルごとに必要部分と不要部分を決定する． 手作業でラベル付けを行った 2583 件（必要語 643 件，不要語 1940 件）のデータを正解データとして用いる．不要語除去性能の評価方法として，評価値には適合率，再現率を用いた．適合率とは，本手法が必要語と判断したものの中で真である形態素の割合である．すなわち商品タイトルの不要語の内，どれだけの不要語を除去することができたかを示した割合である．再現率とは，真の必要語であるうち必要語と分類された割合である．適合率，再現率は式(2)～(3)の通りになる．

表 1 分類法

		分類の結果	
		必要語	不要語
実験データ	必要語	a	b
	不要語	c	d

$$適合率 = \frac{a}{a+c} \quad (2)$$

$$再現率 = \frac{a}{a+b} \quad (3)$$

また、システム不使用の場合と島村らのシステム、本システムを被験者 20 名に使用してもらい、アンケートを行った。

4.2 実験結果および考察

分類結果を表 2 に示した．表 3 は表 2 の分類結果をもとにして算出した適合率，再現率の値を示している．アンケートの結果を表 4 に示した。

表 2 分類結果

		分類の結果	
		必要語	不要語
実験データ	必要語	594	49
	不要語	519	1421

表 3 分類精度

適合率	53.37%
再現率	92.38%

表 4 アンケート結果

	システム不使用		
	収集の容易さ	比較の容易さ	目的に対する満足度
平均値	2.65	2.15	3.60
	島村らのシステム使用		
	収集の容易さ	比較の容易さ	目的に対する満足度
平均値	4.25	3.80	4.00
	本システム使用		
	収集の容易さ	比較の容易さ	目的に対する満足度
平均値	4.10	4.35	4.15

表 3 より，適合率に関しては，約 53%という結果となっており，システム必要語とした中で 53%は正しい必要語であったことになる．これはつまり，商品タイトルから不要語の除去に際して不要語のうち 47%が商品タイトルに残ってしまったことを表している．再現率に関しては，約 92%であり，ほとんどの必要語は誤判別することなく分類を行えている．アンケート結果より、目的に対する満足度としてはシステムの有無でさほど差はないことが分かった。収集比較に関してはシステムにより、収集比較が容易になることが分かった。比較の容易さで島村らのシステムより良い結果であったのは不要語除去の効果であると考えられる。

5 結論

商品タイトルに不要語が混在しているショッピングサイトを含む，複数のショッピングサイトにおいてユーザが興味をもった商品同士のを比較を支援するシステムの構築を行った．

不要語除去性能の評価実験の結果として，必要語と不要語の分類結果から，適合率約 53%，再現率は約 92%という結果が得られた．アンケートではシステムにより、収集比較を容易にするという結果が得られた。

今後の課題としては，不要語除去の性能評価において適合率約 53%と十分な値とは言い難く，不要語除去の性能向上が必要であると考えられる．また、アンケートでは被験者を増やし、より正確な評価が必要であると考えられる。

6 参考文献

[1] 島村祐介, 三末和男, 田中二郎
"ウェブ上の検索システムにおける検索結果の比較支援インタフェース," 情報処理学会第70回大会, 2008
[2] 吉田一星
相対頻度を利用した不要語除去によるテキスト自動分類，情報処理学会第65回全国大会, 2005
[3] 相良毅，牧野俊朗，川口修一，小澤英昭, 喜連川優
“住所情報を用いた店舗名称のクリーニング手法”, データ工学ワークショップ 2006, 2C－o1 (2006)