Wikipedia データベースを用いた スパム判定用ベイジアンフィルタ構成法の研究

福田健太郎[†1]　槇俊孝[†2]　小畑佑介[†1]　若原俊彦[†2]

近年，一方的宣伝・広告メールのようなスパムメール（迷惑メール）がメール全体の約80%を占めるようになっており社会問題になっている．これを解決するため，現在はベイジアンフィルタを用いてメール中に書かれた単語の出現頻度によりフィルタリングする手法が主流となっているが，複合名詞の出現頻度は考慮されていない．そこで本研究では，Wikipediaのデータベース（DB）を活用して複合名詞を同定し，従来システムよりも高精度なスパムメールのフィルタリングシステムを提案し評価する．

A Study on the Bayesian Spam Mail Filter Using Wikipedia Database

KENTARO FUKUDA[†1]　TOSITAKA MAKI [†2]
YUSUKE OBATA[†3]　TOSHIHIKO WAKAHARA[†4]

In recent years, spams are contained in 80% of all email. This is becoming a social problem. Although the Bayesian filter is used for term frequency, compound nouns are not considered. In this study, we propose high accuracy Bayesian spam filtering system by identification of compound nouns using Wikipedia Database.

1. はじめに

電子メールは，インターネット上のコミュニケーション手段の一つであり，その非同期性や記録性，同報性，そして手軽に使用できるという利点から広く普及している．近年では無料で電子メールアカウントが取得でき，メールサービスの管理費用が実質ゼロで利用できる環境も実現している．しかし，電子メールの利便性が高まる一方で，スパムメールが全体の約 80%を占めているという問題がある．スパムメールは，受信者の意志を無視して，無差別かつ大量に送信されるメールである．迷惑メールには以下のようなものがある．

- なりすましメール
  知り合いや有名企業などに似せて，ID やパスワードなどの個人情報を搾取するフィッシングサイトに誘導する．
- 一方的広告メール
  出会い系サイトへの勧誘，購買意欲の無い商品の宣伝等．
- チェーンメール
  連鎖的に不特定多数へ配布するように求める．

これらは近年増加傾向にあり，深刻な被害が出ているケースもあり，社会問題になりつつある[1].

このスパムメール問題を解決するために，様々なフィルタリングが開発され導入している．スパムメール送信者が多く使用するキーワードを登録しておき，そのキーワードを含むメールをブロックするキーワードフィルタ手法，過去に送信されたスパムメール送信者のアドレスや非スパムメール送信者のアドレスを登録しておき，登録されたアドレスのフィルタリングを行うブラックリスト・ホワイトリスト手法，現在までに送信されたメールを解析し，出現したキーワードの機械学習を行いフィルタリングするベイジアンフィルタ手法， From，To，Cc に着目してメールアドレス親密度を算出しメールアドレスのブラックリスト・ホワイトリストを作成しフィルタリングを行う社会ネットワーク手法[2]などのフィルタリング技術が存在する．

電子メールのフィルタリング技術において，性能が良く，一般的に用いられているのはベイジアンフィルタであるが，日本語のスパムメールをフィルタリングする際に問題となる点がある．ベイジアンフィルタは，Paul Graham の研究[3]をきっかけに多く普及すること

†1 福岡工業大学
情報工学部情報通信工学科
Fukuoka Institute of Technology.
Information and Communication Engineering

†2 福岡工業大学大学院
工学研究科情報通信工学専攻
Fukuoka Institute of Technology, Graduate School.
Information and Communication Engineering

となったが，各単語が空白で区切られている英文での使用を前提とし開発された．日本語のように空白で区切られていない文章では単語ごとに抽出することが難しい．このため，本研究では形態素解析エンジンを利用して文章を単語ごとに分割し機械学習し，フィルタリングを行う．しかし，正しく抽出できない場合が多く，2 つ以上の単語からなる複合名詞や未知語の抽出は難しい．複合名詞は単一の名詞では表しきれない固有の概念を表す場合が多く，文章を特徴づけやすい性質がある[4]．したがって，単語だけでは無く複合名詞や未知語を考慮することで比較的高精度にフィルタリングが行えると考えられる．本研究では，ベイジアンフィルタの改善手法として Wikipedia データベースを活用して複合名詞を同定し，また，未知語の固有概念を考慮したベイジアンフィルタ手法を提案しその評価を行う．以下に本論文の構成を示す．第 2 章では本研究の関連研究を示す．第 3 章では従来方式の問題点を示す．第 4 章では提案システムの概要を示す．第 5 章では実験を示し，第 6 章では結果を示す．第 7 章では本研究のまとめを示す．

2. 関連研究

2.1 ベイジアンフィルタ

ベイジアンフィルタとは，ベイズ理論を応用して必要な情報を抽出したり除去するフィルタである．ベイズ理論とは，過去に起きた事象の確率を応用して，未来に起こる事象の確率を予測する理論である．過去に受信したユーザが定義したスパムメール，非スパムメールの機械学習を行い，新たに受信したメールがスパムメールかどうかの確率を計算する．

2.2 Wikipedia

Wikipedia とは, ウィキメディア財団が運営しているインターネット百科事典である．誰でもブラウザを利用して記事内容を編集でき，歴史や数学，科学，社会，テクノロジなどの記事が幅広く網羅されている．平成 26 年 1 月現在，日本語版の総ページ数は 2,483,959 件，記事数は 892,352 件である．Wikipedia は，1 ヶ月以内に日本語版記事の編集を行ったユーザが 18,168 人で編集回数は約 317,000 件であり，リアルタイム性が高く、様々な分野の情報を含む高品質のコーパスとみなすことができる．また Wikipedia の検索機能として，キーワード完全一致の検索だけでなく、「ひらがな」，「カタカナ」，「漢字」などの表記違いに対応したゆらぎ検索機能や、あらかじめ登録されたキーワードと対応づけた検索結果の候補予測の提示を行う機能などがある. Wikipedia のコンテンツは，再配布や再利用のためにデータベースが公開されており，利用者も無償で利用できる．そのため，Wikipedia から言語資源を獲得する研究や辞書作成を行う研究が盛んに行われている．関連研究の中には，「記事ページ「○○一覧」の本文には記事名に属する固有名詞が箇条書きで記述されているものが多い」ということに着目した辞書作成手法の提案 [5]，Wikipedia の記事タイトルを固有名詞に分類[6]，Wikipedia の高密度なリンク構造，コンテンツの網羅性に着目したシソーラス辞書の作成[7]などの例が挙げられる．特に固有名詞，複合名詞の同定に関する研究は盛んに行われている．

2.3 形態素解析

形態素解析とは，人間が日常的に使用している自然言語をコンピュータに処理させるための技術であり，辞書を利用して意味のある単語に区切り，品詞やその他特徴を判定する技術である．本研究では，形態素解析エンジンの 1 つである MeCab を利用する．

3. 従来のベイジアンフィルタの問題点

従来のベイジアンフィルタでは，形態素解析エンジンを利用して対象メール本文を単語ごとに分割して機械学習を行い，フィルタリングを実現している．しかし，形態素解析エンジンは複合名詞や未知語を不適切に処理することがあり，誤った学習結果を導く恐れがある．この問題を解消することでメール本文の意味を考慮した機械学習を行うことができ，比較的高精度なフィルタリングが実現できると考えられる．以下では Wikipedia が配布する日本語版データベースを用いたベイジアンフィルタを提案する．

4. 提案システムの概要

本研究で提案するシステムは, Wikipedia データベースを活用し，複合名詞を同定し，また未知語の出現頻度を用いたベイジアンスパムフィルタを提案する．提案システムは， Paul Graham 方式を用いたベイジアンスパムフィルタを参考にして改良を加えている．Paul Graham 方式ベイジアンフィルタは，機械学習部，判定部の二つのモジュールで構成している．

4.1 機械学習部

ベイジアンフィルタにおける機械学習とは，受信したメールをスパムメールか非スパムメールかの確率を計算する処理のために各メールの特徴を抽出する前処理である．抽出を行うためには受信者が，スパムメールと非スパムメールを手動で定義する必要がある．次に定義したスパムメールと非スパムメールを 2 つの集合とし，メール本文の形態素解析を行う．英文の場合，単語間が空白で区切られているため，基本的には単語を単位として分割する．日本語の場合では，空白で単語間が区切ら

れていないため，単語の抽出方法が大きな課題となっている．単語ごとの分割処理では， yahoo! Japan の形態素解析を使用する手法, MeCab や Juman などの形態素解析エンジンを利用する手法などがある．本研究における機械学習では，MeCab を利用して単語の抽出を行い，単語ごとの出現頻度と合わせてデータベースに格納する．なお，データベースはスパムメールと非スパムメールでクラス分けており，それぞれを対応させている．

4.2 対象メールのスパムメール確率の計算

本研究では，対象メールを Paul Graham 方式によるベイジアンフィルタによってスパムメールの確率を算出する．対象メールがスパムメールと判定する確率を計算するために，対象メールの本文を単語ごとに分割し，単語ごとのスパムメールである確率を計算する必要がある．データベースに登録された単語ごとの出現頻度をもとにスパムメールである確率を計算する．

データベースの情報をもとに，単語 **wi** がスパムメールである確率は次式で表わされる．

200

$$P(wi) = \frac{\left(\frac{bi}{nbad}\right)}{a\left(\frac{gi}{ngood}\right) + \left(\frac{bi}{nbad}\right)} \quad \cdots (1)$$

a ： バイアス(2 が使われることが多い)

gi ： wiが非スパムメールに登場した回数

bi ： wiがスパムメールに登場した回数

スパムメールと判定する確率は，0 から 1 の値域で表され，1 に近いほどスパムメールである確率が高い．また， 0 に近いほど非スパムメールであることを表している．式(1)から求められた単語ごとのスパムメールである確率$P(wi)$を用いて，対象メール D がスパムメールである確率は式(2)で表わされる．

$$P(D) = \frac{\prod_{i=1}^{n} p(wi)}{\prod_{i=1}^{n} p(wi) + \prod_{i=1}^{n}(1 - p(wi))} \quad \cdots (2)$$

$P(D)$ ： 対象メール D がスパムメールである確率

$P(wi)$ ： 単語ごとのスパムメールである確率

式(2)から求められた単語ごとのスパムメールである確率$P(D)$が，受信者のあらかじめ設定した閾値を下回った場合は非スパムメールと判定する．

4.3 提案システムの処理概要

ベイジアンフィルタの機械学習部分における単語の登録の際に, Wikipedia データベースの記事ページ情報を利用した手法を提案する．

本研究では， 前述のように Paul Graham 方式を用いたベイジアンスパムフィルタ，形態素解析エンジンには MeCab, 複合名詞の同定，未知語の検出には，Wikipedia データベースをそれぞれ利用する．見出し語辞書内には複合名詞でないものも多く含まれているため MeCab 解析結果の品詞情報を利用し複合名詞を適切に同定できるようにしている．対象メール本文を一行ごとに読み込み，形態素解析エンジンを利用し，形態素ごとに分解し品詞や内容を判別し，連続した名詞のみを抽出する．抽出方法を図 1 に示す．

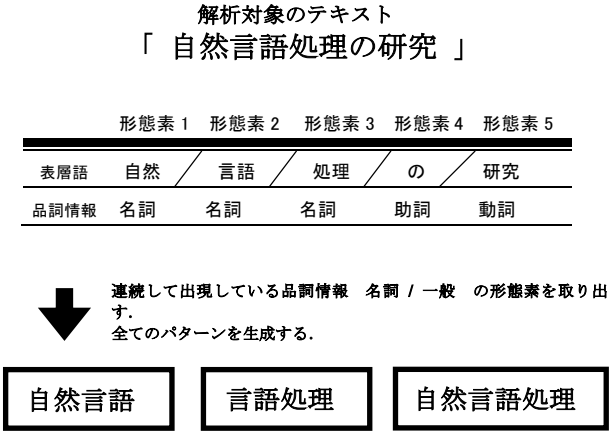

図 1 複合名詞候補の組み合わせ

抽出した連続した名詞の組み合わせを複合名詞候補とし，Wikipedia データベースと比較する．対象メールから複合名詞を抽出するために，Wikipedia の記載ページタイトル(見出し語)，記載ページに関する関連キーワードから検索を行う．記載ページタイトルは，Wikipedia データベース内の page テーブルから抽出を行う．抽出した単語によっては，検索結果が複数個存在する記載ページタイトルも存在する．多くのページとリンクしており一般ユーザが多く閲覧する最適な記載ページを判別するために，Wikipedia データベースの名前空間(name_space)を利用した．名前空間とは，Wikipedia 上の各ページ属性である．見出し語の検索には，通常の百科事典記事に使用されている標準名前空間に属する記載ページタイトルを使用する．標準名前空間に属する記載ページタイトルを見出し語辞書として複合名詞の検索を行う．見出し語辞書に関する

情報を図 2 に示す.

カラム名	型名	コメント
Page id	int	記載ページのインデックス
Page Namespace	int	ウィキペディア上の各ページが属する分野
Page_title	char	記載ページのタイトル名

複合名詞候補を補抽出しの文字列を見出し語辞書のPage_titleと比較し一致した場合，複合名詞とみなす.

Page_id を用いて WikiPageLink から「自然言語処理」の関連情報取得

複合名詞	自然言語処理
カテゴリ	理論言語学
関連キーワード	形態素解析
	構文解析
	言語獲得
	データベース
	⋮

図 2 見出し語辞書を用いた複合名詞の抽出，情報取得

形態素解析エンジンを使用し分割した単語の出現頻度と，見出し語辞書を用いて抽出した複合名詞から得られた関連キーワード，そしてその出現頻度を利用して非スパムメール，スパムメールごとに非スパムデータベース，スパムメールデータベースに格納する．ベイジアンフィルタの機械学習のシステムフローを図 3 に示す.

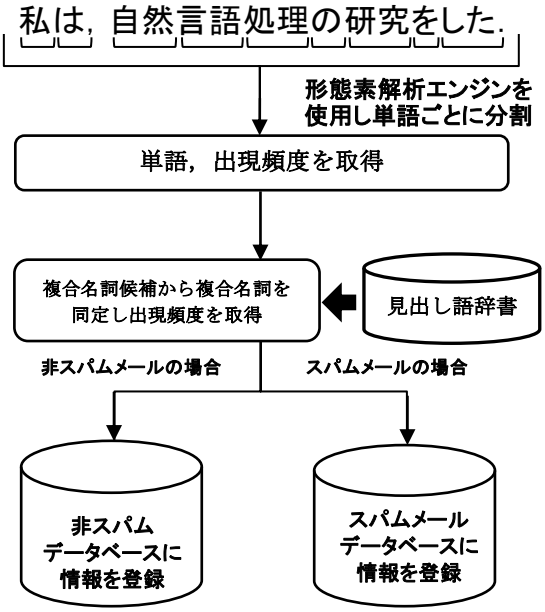

図 3 ベイジアンフィルタにおける機械学習処理

機械学習によって得られた 2 つのデータベース情報を利用し，ベイジアンフィルタの判定処理を行う．フィルタリング対象メール本文は，Mecab を用いて単語に分割する．分割した単語，および複合名詞ごとに式(1)を用いてスパムメールの確率を計算する．分割した単語，複合名詞ごとのスパムメール確率と式(2)を用いて，スパムメール確率の計算を行い，管理者の設定した閾値を上回った場合スパムメールと判断し，下回った場合は非スパムメールと判断する．ベイジアンフィルタのフィルタリングのシステムフローを図 4 に示す.

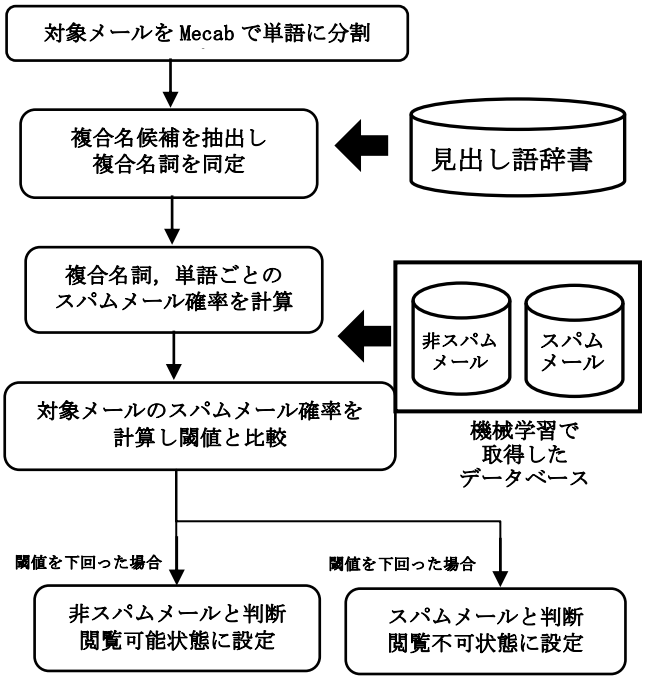

図 4 ベイジアンフィルタにおけるフィルタリング処理

提案システムでは，名詞，複合名詞と判定された単語は従来のベイジアンフィルタと同様に式(1)でスパム確率を求める．複合名詞を構成する単語のスパムメール確率は対象メールのスパムメール確率を求める際に除外する．フィルタリング時の，複合名詞同定処理のシステムフローを図 5 に示す.

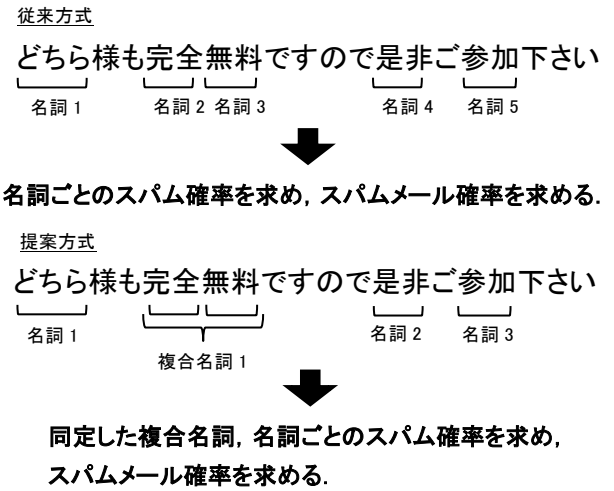

図 5 フィルタリング時の複合名詞同定処理

以上の手法を用いて複合名詞の出現頻度を用いたベイジアンスパムフィルタを実験により評価する.

5. 実験による提案手法の評価

提案したベイジアンスパムフィルタの性能評価を行う. 実験条件を以下のように設定する.

5.1 評価対象

実験対象として使用するメールは, 2012 年 4 月から 2013 年 4 月の 1 年間に研究室のある特定の一個人が受信したものであり，Yahoo!メールおよび Gmail から抽出している. 実験に使用する学習用と評価用メールの内訳を表 2 に示す.

表 2 学習・評価用メール

	種類	件数(件)
学習用メール	非スパムメール	700
	スパムメール	700
評価用メール	非スパムメール	1,133
	スパムメール	261

5.2 Wikipedia，見出し語辞書

見出し語辞書に利用する Wikipedia データベースとしては, 2014 年 1 月 8 日のデータベースダンプしたものを使用する. 使用したデータは，ページ情報(Page_id,Title)が登録されている page.sql，ページ間のリンク情報が登録されている agelinks.sql，ゆらぎ検索機能で使用する redirect.sql，カテゴリ情報を登録してある category.sql を用いて複合名詞の検出を行っている. データベースの詳細を表 3 に示す.

種類	件数(件)
記載ページ	891,442
総ページ数	2,480,555

表 3 Wikipedia データベース詳細

5.3 評価対象範囲

評価対象範囲は，テキスト形式で保存されているメール本文のみである. ヘッダー情報(タイトル，送信元メールアドレス，受信日時)の評価は行わず，データベースへの登録のみ行う. テキスト形式で保存されているメールは，Thunderbird を用いて Yahoo!メール，Gmail から収集したものである. Thunderbird は，Mozilla Foundation が開発を行っている電子メールクライアントである. テキスト形式で保存されたメールの構成を図 6 に示す.

Subject: "タイトル"

From: "送信元メールアドレス"

Date: 0000/00/00 00:00:00 "送信日時"

To: "宛先メールアドレス"

Reply-to: 返信先メールアドレス

本 文

図 6 対象メールの構成

6. 実験結果

従来方式と提案方式を誤遮断率(FPR : False-Positive Rate = 非スパムメールがスパムメールとして遮断されてしまうこと), 誤通過率(FNR : False-Negative Rate = スパムメールが非スパムメールとみなされフィルタを通過してしまう確率), 誤判定率(ER : Error Rate = スパムメールであるかどうかを誤判定されてしまう確率，誤通過率，誤遮断率してしまう確率)を評価した. 従来手法と提案手法の実験結果を表 4 に示す. ベイジアンフィルタ単独の場合に比べ FPR は同じであったが，ER，FNR は 5%台から 4%台と 1%以上低下した. またスパムメールの検出率は 95.91%となり 1%以上向上した. 実験で定義したスパムメールでは，不正 B-CAS カードの広告メールが多く届いている. その表現の中で「どのチャンネルも完全無料で視聴可能!!」など特徴的な複合名詞が高頻度で出現する. しかし,「完全」と「無料」という単語は，非スパムメールの中に，出現することが多くスパムである確率は，少なくなってしまう. この提案手法により，単語だけでは特徴分析が難しいスパムメールも複合名詞の出現頻度を考慮することで特徴的な複合名詞の確率も考慮できるようになった. 例としてスパムメールデータベースにて出現頻度の多い複合名詞と構成する単語のスパムメールである確率との比較を図 7 に示す.

単語の場合のスパム確率

完全	無料
スパム確率	スパム確率
0.621	0.304

複合名詞の場合のスパム確率

完全無料
スパム確率
0.98165

図 7 複合名詞と単語のスパムメール確率の比較

方式に比べ良好な結果を得られることを確認でき，複合名詞を考慮したフィルタリングシステムが有効であることを実証できた．しかし，提案手法では，品詞情報に動詞，助詞，助動詞を含む単語で構成される固有名詞などの検出は行えない．固有名詞にも，独特の固有概念を表す場合が多く，文章を特徴づけやすくスパムメール確率を求める際の指標の一つとして利用できると考えられる．これを解決するため，見出し語を用いて全文検索を行う方法があるが，見出し語辞書に用いる Wikipedia データベースには膨大な量のデータが格納されており，検索を行うために膨大な時間が必要となる．

表 4 従来手法と提案手法の比較

	成功率	誤遮断率	誤通過率	誤判定率
従来手法	94.35%	0.14%	5.37%	5.52%
	1317/1396	2/1396	75/1396	77/1396
提案手法	95.91%	0.14%	4.08%	4.20%
	1338/1396	2/1396	56/1396	58/1396

実験で同定すべきと判定できた複合名詞の各メール出現頻度分布を図 8 に示す．

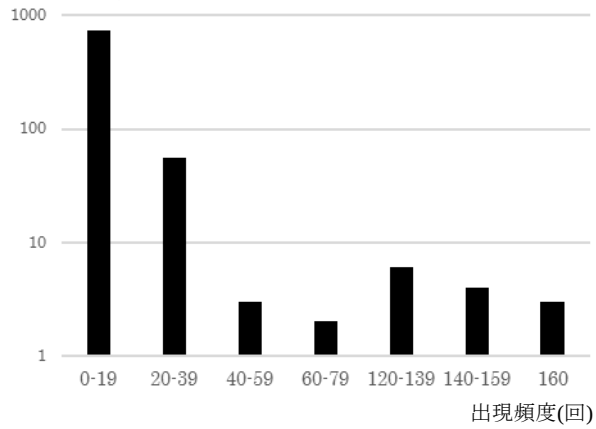

図8 複合名詞と単語のスパムメール確率の比較

7. まとめ

本研究では，従来のメールフィルタリングでは検知の難しい複合名詞の同定処理を導入したベイジアンフィルタシステムを提案した．これにより，スパム確率の高い複合名詞を分割された単語の出現頻度でスパム確率を求めず，複合名詞を同定してスパム確率を算出することが可能となった．複合名詞は固有の概念を表す場合が多く，文章を特徴づけやすい．提案手法をシステムに実装し，実際に受信したメールを用いて実験を行った結果，提案システムは従来

今後は，新たな品詞の単語を考慮した固有名詞の同定処理を行い，より上位の語辞書の最適化手法を検討する．

文 献

[1]神谷 造, 柴田 賢介, 佐野 和利, 荒金 陽介, 塩野入 理, 金井 淳史 ”スパムメールの外見的特徴についての一考察” 電子情報通信学会技術研究報告 ISEC，情報セキュリティ 107(141) 123-127 2007-07-13

[2]大福 泰樹，松浦 幹太 ”ベイジアンフィルタと社会ネットワーク手法を統合したスパムメールフィルタリングとその最適統合法情報処理学会論文誌 情報処理学会論文誌 47(8), 2548-2555, 2006-08-15 一般社団法人情報処理学会

[3] Paul Graham “Hackers & Painters: Big Ideas from the Computer Age” O'Reilly Media; 1 版 (2008/7/14)

[4] 近藤智司，西岡悠，稲葉宏幸”品詞情報に着目した日本語スパムメールフィルタリングの提案” 電気学会電子・情報・システム部門大会講演論文集 2007 MC2-6 2007/09/04

[5] 田村 直之 伊藤 直之 西川 侑吾 中川 修 新堀 英二 ”Wikipedia から作成した辞書によるブログのカテゴリ分類” FIT2009 E-010

[6]杉原 大悟，増市 博，梅元 宏，鷹合 基之 ”Wikipedia カテゴリ階層構造の固有名詞分類実験における効果” 情報処理学会研究報告 自然言語処理研究会報告 2009, 57-64, 2009-01-15 一般社団法人情報処理学会

[7]中山 浩太郎，原隆浩，西尾 章治郎 ”Wikipedia マイニングによるシソーラス辞書の構築手法” 情報処理学会論文誌 47(10), 2917-2928, 2006-10-15 一般社団法人情報処理学会