社団法人 情報処理学会　研究報告
IPSJ SIG Technical Report

2004－NL－159　(3)
2004／1／13

ソフトウエア製品ニュースからの開発傾向の抽出

徳永秀和†　　天雲 勇作†　　青江順一‡

近年，インターネット上に多量の新製品ニュース記事が存在する。これらの記事の有効利用のひとつとして，ソフトウエア製品ニュース記事から最近の開発傾向を探るシステムの開発を行うこととした。そのためには，製品の機能を示すキーワードを抽出することが必要となる。この抽出を，サポートベクターマシンで行った。さらに，抽出もれを補うため，全文検索と重み付け統計を取ることによってソフトウエア製品の開発傾向が得られることを報告する。

Extraction of The Development Tendency from Software Product News Article

Hidekazu TOKUNAGA†　　Yusaku TENKUMO†　　Jun-ichi AOE‡

At present, a lot of new product news article exist on the Internet. So, it is effective to develop the system which explores the latest development tendency from a software product news article. It is necessary to extract the keyword which expressing the function of a product. We performed keyword extraction using the support vector machine. Furthermore, in order to compensate the omission in extraction, the full-text search and statistics of weighted keywords was performed. Consequently, we report that the development tendency of a software product was acquired.

１．はじめに

現在，コンピュータの急速な進化とネットワーク環境の整備により，一般人でもwwwにネットワーク接続しているパソコンさえあれば多くの情報を閲覧可能である。しかし，多くの情報が得られる反面その中から目的の情報を探す労力がより多く必要になる。例えば，多くのソフトウエア製品ニュースから最近の開発傾向を知ろうとすると，一つ一つのニュース文を見ていく必要があり，ユーザ側に大変な労力を強いる。そこで，ニュース記事より製品の機能的特徴を抽出し，開発傾向をユーザに分かりやすく提示するシステムを開発することが有用である。

本論文ではセキュリティ分野に限って、ＳＶＭ（サポートベクターマシン）を用いて，製品の機能的特徴を示すキーワードをラベル付けしたニュース記事を学習させることによって，未知のニュース記事からキーワードを抽出する。得られたキ

†高松工業高等専門学校
Takamatsu National College of Technology
‡徳島大学知能情報工学科
Dept. of Information Science & Intelligent Systems, University of Tokushima

ーワードを全文検索と重み付け統計を用いて，開発傾向を提示するシステムに関する研究結果を示す。

２．ＳＶＭ（サポートベクターマシン）

サポートベクターマシン[1]は，特にパターン認識の能力について，現在知られている中でもっとも優秀な学習モデルのひとつである。ＳＶＭというのは「線形しきい素子」というモデルに「マージン最大化」という基準で学習させ、更にカーネルトリックという工夫を加えて性能を上げたものである。

２．１　線形しきい素子

ＳＶＭの基礎となる線形しきい素子は，ニューロンを非常に単純化したモデルで，入力ベクトル $\mathbf{x}$ に対して，２値$(-1, 1)$の出力 y を識別関数（１）に従って出力する。

$$y = \mathrm{sgn}[\sum_{i=1}^{n} < w, x > -h] \quad \cdots\cdots (1)$$

$\langle \mathbf{w}, \mathbf{x} \rangle$はベクトルの内積をあらわしており，

$\boldsymbol{w} = (w_1, w_2, \cdots, w_d) \in \boldsymbol{R}^d, h \in \boldsymbol{R}$ である。

また，$\mathrm{sgn}[u]$は$u > 0$で１，$u \leq 0$で－１をとる符号関数である。このモデルの働きを幾何学的に説明すると，入力空間である $\boldsymbol{R}^d$ を $< \boldsymbol{w}, \boldsymbol{x} > -h = 0$ で定義される超平面で二つにわけ，一方に１をもう一方に－１を対応させることにあたる。線形しきい素子の学習は，与えられた例題に対して，$\boldsymbol{w}$ やh というパラメータを調整することにより行われる。

２．２　マージン最大化

すべてのサンプルに対して正しい出力を出す超平面は一意的ではない，そこでそのような超平面のうち汎化能力の点において最適なものがあるはずである。直感的に考えると訓練パターンすれすれを通る超平面よりは多少余裕をもって分ける超平面のほうが良いと思われる。ＳＶＭではその余裕をマージンという量で測り，マージンをできるだけ大きくするような超平面を求める。マージンは超平面と一番近くにある訓練パターンとの距離の最大値である。

３．システム概要

３．１　システム構成

まず学習用にWeb上から取得したセキュリティ関連のニュース記事を形態素解析し，形態素単位で抽出（ｙ＝１）と非抽出（ｙ＝－１）のラベル付けを人手で行い学習用ファイルを作る。そして次節で示す特徴ベクトルを作成し，ＳＶＭで学習させる。その学習ずみのＳＶＭを用いて未知のニュース記事からキーワードを抽出する。そして抽出もれを補うため，得られたキーワードを未知の記事全体に渡って全文検索を行い，重み付け統計によって開発傾向を示す。

３．２　特徴ベクトルの各要素

本研究での特徴ベクトルは，山田らの論文[2]を参考にし，必要最小限の要素をいくつか試した結果一番精度が良い傾向にあった次のようなものとした。

１．対象単語の品詞

２．対象単語の隣にある左右の単語の品詞

３．対象単語の文字種（英字，カタカナ，平仮名，漢字，数字，記号）

４．対象単語の隣にある左右の単語の文字種

５．対象単語が記事全体の何文目か

６．対象単語の文の手前部分に特定単語（ソフト，ソフトウェア，ツール，システム，パッケージ）がいくつ存在するか

７．対象単語の文の後ろ部分に特定単語（上記）ががいくつ存在するか

３．３　特徴ベクトルの例

表１は元記事の例として取り上げたニュース記事である。このニュースはDELETE MASTERという個人情報保護ツールについての記事である。

まずこの元記事を，文章の各単語の形態素を得るために，形態素解析ソフトchasen[3]にかけて各単語の形態素を得る。得られた結果は表２の通りで，この出力結果は左から単語そのもの，単語の

表１　元記事の例

メディアヴィジョン,個人情報保護ツール（DELETE MASTER）を発売

2003年7月29日

（株）メディアヴィジョンは２９日、個人情報保護ツール『DELETE MASTER（デリートマスター）』、『DELETE MASTER Personal（デリートマスター　パーソナル）』を

・
・

表２　Chasenの処理結果

メディアヴィジョン．メディアヴィジョン．メディアヴィジョン．１７００

、．、．、．７９００

個人．コジン．個人．２００
・
・

表３　ラベル付け済みのデータ

×．メディアヴィジョン．メディアヴィジョン．メディアヴィジョン．１７００

×．、．、．、．７９００

○．個人．コジン．個人．２００
・
・

読み，単語の原型，最後に品詞とその活用を数値で表したものである。

そして，この段階で学習用データには人間の手でラベル付けを行う。この例では「個人情報保護ツール」というキーワードに正解の○をつけ，それ以外は×をつける。そのようにして正解のラベル付けを施したものが表３である。

次に,表 ３のデータを読み込み特徴ベクトルを作るプログラムを実行し，特徴ベクトルを作成したものが表５である。これの各要素を説明すると，左からラベル（1.0が抽出対象、0.0がそれ以外である。），残りは要素のインデックス（表 ４）（そのデータがどの要素のものか示すもの）と値が交互に入っている。表 ５で一番上の行で例を挙げると左から最初はラベルで次が０なので表 ４ のインデックス関連表より左側の単語の品詞だが単語自体存在しないので０とする，次は３だがこれも左側に単語が存在しないことから文字種も０，次は１で対象単語の品詞を表すここでは chasen の出力より未知語などを表す１７という数値が入っている。次は４で単語の文字種の値２（カタカナ）が次に入っている。このように表 ４を参照しながら表 ５ を見ることで特徴ベクトルがどのようなものか理解できる。

表 ４　　インデックスの関連表

0	左隣の単語の品詞
1	対象単語の品詞
2	右隣の単語の品詞
3	左隣の単語の文字種
4	対象単語の文字種
5	右隣の単語の文字種
6	その単語が存在する文において、左側に存在する特定の単語の数
7	その単語が存在する文が何文目に存在するか
8	その単語が存在する文において，右側に存在する特定の単語の数

表 ５　　学習用特徴ベクトル

0.0 0 0 3 0 1 17 4 4 2 79 5 5 …
0.0 0 17 3 4 1 79 4 5 2 2 5 6 …
1.0 0 79 3 5 1 2 4 6 2 2 5 6 …
・
・

３.４　学習と抽出

表 ５の特徴ベクトルのパターンをＳＶＭに学習させる。そして学習済みＳＶＭを用いて未知の記事に対して，同じように各単語の特徴ベクトル（ただし，表 ５の左端の正解のラベルはない）を作成し各単語をラベル付けする。最後にＳＶＭの出力結果より正解のクラスラベルがついている単語を取り出すことで抽出対象を取り出す。

３．５　全文検索と重み付け統計

ＳＶＭでのキーワード抽出には抽出漏れが多いためその抽出漏れを防ぐために全文検索と重み付

け統計という手法を用いる。ＳＶＭでの抽出において同じ単語がキーワードである場合でも記事によって抽出ができたり，キーワードの一部分しか抽出できないことが頻繁にある。全文検索はそのような場合を防ぐために得られたキーワードを記事全体に渡って全文検索にかけ該当する単語を拾い出すことによりＳＶＭで抽出できなかった単語を抽出しようという試みである。だが，全文検索を用いることによって問題も発生する。ＳＶＭで望むべきものではないキーワードを抽出した場合にそのキーワードをも全文検索にかけてしまうことになる。そこで，抽出される不要キーワードが１単語だけである（複合語でない）ことが多いのと，キーワードが１単語で構成されていることがほとんどないことを利用し，記事をＳＶＭで抽出された単語のみ表示し，その中で複合語となっている場合のみその複合語をキーワードとして抽出することで解決する。

だが、２単語以上で構成される不要なキーワードの抽出や，本来その記事の主題ではない部分に存在したキーワードの抽出などの問題が残る。そこで，なるべく影響を押さえる対策として取り入れたのが重み付け統計という手法である。この手法は抽出したキーワードそれぞれに記事をヒントに重みをつけるのである。ただ、この手法を有効に使うには，必要なキーワードの重みを増やし不要なキーワードの重みを減らすようなルールをうまく考える必要がある。これにより統計の結果も大きく変わってくる。本研究では次のようなルールを用いる事によって重み付け統計を行う。

［以下の場合重みを増やす］

・そのキーワードがある文に発売、発表などの比較的ソフトウェアの特徴を表す語と同時に存在しやすい単語がある場合。

・キーワードが記事の始めのほうに存在する場合。

［以下の場合重みを減らす］

・キーワードが一つの記事に複数含まれる場合。

・キーワードが記事の終わりのほうに存在する場合。

４．実験

４.１　実験方法

［実験１］

ASCII24.com よりセキュリティ関係のソフトウェアニュース記事を３０記事（２３種類のキーワード，キーワード出現回数９７回）を集めて，その記事を形態素解析にかけ、単語単位でラベル付けする。ここで正解ラベル付けする単語は，その記事が何に関してのものであるかというもの（例えばウィルス対策ソフト，個人情報保護ツール，総合パッケージなどその記事のソフトウェアがどの種類のソフトウェアであるか判断可能であるもの）である。そのラベル付け済みの３０記事分の特徴ベクトルを作成しＳＶＭに学習させる。次に，学習済みのＳＶＭより未知の５０記事（キーワード種類３９種類、キーワード出現回数１７９回）からキーワード抽出がどの程度可能であるかを調べる実験を行う。

［実験２］

実験１の結果を元に全文検索を行う。実験１でのＳＶＭにより抽出した全てのキーワードを単語に分別する。例えば「ウィルスソフト」を「ウィルス」と「ソフト」に分別し，リストに保存する。そしてリストに存在する単語だけを元記事において表示する。図１に一文だけでこの全文検索を行うとした場合の手順を示す。これは１文だけの場合であるがこれを全記事に渡って行う。この実験ではこの全文検索を使ってキーワードがどの程度取得できるかを検証する。

［実験３］

実験２の全文検索のデータの統計とそれを重み付け統計したものによって，全体の傾向の比較をすることにより重み付け統計の有効性について検証する。ただし，統計は人手によるものであり，類似語（セキュリティソフトとセキュリティツールなど）を一つにまとめたり，実験２で完全に抽出されているキーワード内で部分的なキーワードを取り除いている。例えば「ウィルス対策ソフト」

＜元記事＞
「▲▲がウィルス対策ソフト〇〇というソフトを発売した。」
↓
ＳＶＭでキーワード「ウィルス対策」，「ソフト」が抽出されたとする。
↓
リストには「ウィルス」，「対策」，「ソフト」が入る。
↓
元記事においてリストにあるワードだけ掲示する。
「　　ウィルス対策ソフト　　ソフト　　」
↓
この中で２単語以上で構成されているワードを拾いだす。ここでは「ウィルス対策ソフト」が拾いだされる。

図１．全文検索の例

が完全なキーワードとすると，「ウィルス対策」，「対策ソフト」は不要なキーワードとして排除している。ただし，記事全体においてその完全キーワード（例では「ウィルス対策ソフト」）の出現数が多い場合に限る。理由としては「管理ソフト」などキーワードでは，詳しく分けると多種なキーワード（個人情報管理ソフト，資産管理ソフト，etc・・・）がある場合別々ではランキングのかなり低い所に位置づけされてしまうが「管理ソフト」とするとランキングの上位にランク付けされ，傾向の把握にも有益な情報であるような場合があるためである。

４．２　実験結果

４．２．１　実験１

学習済みのＳＶＭで未知の５０記事（３９種類のキーワード，キーワード出現回数１７９回）からキーワード抽出を行った結果を表 6 に示す。表 6 中で部分的，完全の意味は、例えば「ウイルス対策ソフト」の場合「ウイルス」「対策」「ソフト」の３つの単語で構成されておりこれを全部抽出できれば完全，どれかが抽出できていれば部分という意味である。

表 ６　実験１の結果

完全に抽出されたキーワード種類	6 種類(6/39，15%)
部分的にしか抽出できなかった キーワード種類	25 種類 (25/39，79%)
完全にキーワードが抽出された個数	15 個(15/179，8%)
部分的にキーワードが 抽出された個数	67 個 (67/179，４６%)
抽出された不要ワードの 種類と個数	８種類 13 個
抽出元の記事の特徴を表してなく キーワード種類には存在するもの	９種類 ２９個

４．２．２　実験２

実験２の結果を表 7 に示す。これを見ると完全抽出キーワード種類がかなり増えておりＳＶＭで完全に抽出できていなかった部分を補っているのが見て取れる。

表 7　実験２の結果

完全抽出キーワード種類	２２種類
部分のみ抽出キーワード種類	９種類
抽出された不要ワードの種類	３種類
抽出された不要ワード出現回数	２３回

４．２．３　実験３

実験３での結果、統計のランキングを取ったものを表 ８に示す。A,B,C はそれぞれ自分での統計，全文検索のみを用いた統計，実験３の重み付けと全文検索を用いた統計の結果である。

４．３　考察

実験１の結果を見ると，キーワードを完全に抽出できているのは僅か１０%程度である。ただ部分的に抽出されているキーワードは多く，しかも部分的にとられているのは毎回同じ部分ではなく違う部分である場合が多くあり，部分的なものを

表 ８　抽出、検索結果の統計（上位６ワード）

	A	*B*	*C*
	管理ソフト	ウィルス対策	管理ソフト
	ウィルス対策	管理ソフト	ウィルス対策
	暗号化ソフト	電子メール	セキュリティ
	セキュリティ	セキュリティ	電子メール
	バックアップ	サーバー用ソフ	バックアップ
少	メールソフト	バックアップ	サーバー用ソフト

Ａ・・・自分で統計した結果

Ｂ・・・全文検索のみの統計結果

Ｃ・・・全文検索した上で重み付け統計結果

組み合わせると完全になるものが比較的多く見られる。他にもその記事の特徴を表していないが他の記事でのキーワードとなるものを抽出しているものも見受けられる。これは，後の加工の仕方によっては使えるデータが抽出されたとも言えるだろう。また、不要なキーワードが抽出されてしまっているが，ほとんどが１単語で構成されているものであり，全文検索の過程で排除されると考えられる。

次に実験２の結果について考察してみる。実験１の抽出データの傾向からも分かるように表.８を見ると明らかに完全キーワードが増えているのがわかる。実験１の考察で述べた事が原因と考えられる。ただ，不要キーワードの種類が少ない割に出現回数が多いのが問題である。このようなデータは後の統計時に大きく反映されることが予想できる。このように全文検索はＳＶＭでの抽出を補っているが新たな問題点も生じる事がある。

最後に，実験３について考察する。まず表 ９を見て一番に目につくのが「暗号化ソフト」がない点である。これはＳＶＭによる抽出の時点で抽出できておらず後の処理ではどうしようもないのが現状である。やはり，ＳＶＭ抽出自体の精度を上げる必要性があるだろう。次に目につくのが電子メールがランキングに入っている点である。これはあまり望ましくないキーワードであるのだが，キーキーワードの中に電子メールソフト等があったため全文検索でウイルス対策ソフトなどの記事に多く含まれる電子メールというキーワードが多く検出されたのが原因である。しかし，重み付け統計されたＣを見みると電子メールのランクが落ちており幾分実際の傾向に近いものとなっている。これは記事の後のほうの説明で多くでてきている電子メールというキーワードの重みが下がったためだと思われる。重み付け統計が有効に働いた結果だと言える。

５．結論

現在の問題点としていくつかある。まず，特定の分野でしか実験を行っていないことから，様々な分野の場合に通じるか分からないというのが現状である。他に統計の際に類似キーワード（例えば「セキュリティソフト」と「セキュリティツール」など）の場合をどうプログラムで判断させるかということが全文検索と統計の自動化を行う上で課題である。

参考文献

[1] Vapnik,V.N.:Statistical Learning Theory,A Wiley-Interscience Publication，1998

[2] 山田　寛康，工藤　拓，松本　祐治，”Support Vector Machine を用いた日本語固有表現抽出”，情報処理学会論文誌，Vol.43，No1，PP44-53，2002

[3] http://chasen.aist-nara.ac.jp/