ネットワークの情報を利用した概念獲得支援

Concept Acquistition Support with Information on Network

黒田 直志　　大杉 英一*　　岩爪 道昭

Tadashi Kuroda　　Eiichi Ohsugi*　　Michiaki Iwazume

武田 英明　　西田 豊明

Hideaki Takeda　　Toyoaki Nishida

奈良先端科学技術大学院大学 情報科学研究科

Graduate School of Information Science, Nara Institute of Science and Technology

Abstract: Recently, the number and variety of information resources on the Internet have been increasing repidly. As more information become available on the Internet, it becomes increasingly difficult for us to get information we need. In this paper, we propose a supporting method of concept acquistition. We chose partial terms, which appear on related pages but not appear on unrelated pages, as indicator. To examine our approach, we reclassified data using feature vectors whose elements were acqired by the method. The result shows that this method is better than a conventional method in classification of data and effective in concept acquistition.

1　はじめに

近年、インターネットなどの情報ネットワークを利用する人が急激な勢いで増加している。これに伴い、これらの情報ネットワーク上で利用できる情報の量は、爆発的に増加し、その種類も多岐に及ぶようになった。このことにより、個人、あるいは企業などの情報ネットワーク利用者は、多種多様な情報を、ネットワークを通して得ることが可能となった。しかしその一方で、情報の量が爆発的に増加したことにより、ユーザが人手によって情報の収集や分類を行うことが難しくなってきている。このため、ユーザが本当に必要としている情報を得るのが難しくなっており、また、得るまでにかなりの労力を要するのが現状であり、ユーザが情報ネットワークを利用しようという意欲を失う原因になりかねない。そのため、自動的にネットワーク上の情報を収集、分類してユーザに提供できるようなシステムの開発が望まれている。

ネットワーク上の情報を自動的に収集、分類する方法の一つとして、我々はオントロジーを利用する方法を提案した [3]。抽出したい情報の概念体系を記述することにより、オントロジーを用いた情報抽出が可能となる。しかし、現在このような概念体系の記述は手作業で行われており、大変な手間と時間が必要となるうえに、構築の際の間違いや見落しの可能性もある。これらを解消するためには、概念獲得の支援が必要となってくる。

そこで、本研究では、ある概念に関する情報をネットワーク上から取り出し、そこに出現する単語の重要度を「情報量」を用いて調べ、その概念に対する情報価値の高い単語を抽出した。また、それらを用いて情報の再分類を行った。再分類の際にはデータ分類のための特徴ベクトルを用いた。その結果、情報量の計算により抽出された単語が概念に大いに関連があり、概念の獲得の際の支援になることがわかった。

連絡先: 奈良先端科学技術大学院大学，情報科学研究科，
情報処理学専攻，知能情報処理学講座，
〒 630-01 生駒市高山町 8916-5，
phone:07437-2-5265，fax:07437-2-5269，
e-mail: tadash-k@is.aist-nara.ac.jp

*現在 NTT に所属

2　情報量を用いた単語抽出

パソコン関連製品の情報を対象として、WWW 上の情報から概念に関する単語の自動抽出を行った。その過程は次のようになる。

1. WWW ページを収集する。
2. 各 WWW ページについての単語リストを生成する。
3. 手動で各 WWW ページをある概念に関連するか否かを決定する。
4. 概念に対する各単語の重要度を決定する。
5. 概念に対する重要度が高い単語を上位 20 個抽出する。

ある概念に関連するページの多くに出現するが、その概念には関連しないページにはあまり出現しない、つまりその概念に関連するページに偏って出現する単語が、その概念に対する重要度が高い単語である、と考えられる。そこで、ある概念に対する各単語の重要度を決定する際に、情報量という値を用いる。ここでの情報量は、ある単語が、収集された WWW ページにどのような分布で出現するのかを示しており、次式により定義される [2]。

$$\begin{aligned} E(W,S) &= -P(W=present)I(S_{W=present}) \\ &\quad -P(W=absent)I(S_{W=absent}) \end{aligned}$$

但し、

$$I(S) = \sum_{C\in\{hot,cold\}} -p(S_c)\log_2(p(S_c))$$

である。

概念に関連するページ (以下、 *hot* なページ)、または概念に関連しないページ (以下、 *cold* なページ) のどちらかに単語が偏っている場合には、情報量の値は小さくなる。 *cold* なページ群には、他の概念の *hot* なページなどがまべんなく含まれているため、 *cold* なページに単語が偏ることはない、という仮定の下で、情報量の値が小さい単語は、その概念に対する情報価値が高い、と判断した。実際に、パソコン関連製品の情報から、情報価値の高い単語を 20 個抽出した結果、その中には *cold* なページに偏って出現する単語は含まれなかった。

また、ある概念に重要な単語を抽出する方法としては、今回使用した情報量の値を基にする方法の他にも、 hot なページに含まれる総数が多いものから順に抽出する方法や、その単語が出現する hot なページの数が多いものから順に抽出する方法などが考えられる。それぞれの方法を用いて、 PRINTER という概念に関する重要単語を上位 20 個抽出した結果、表 1 のようになった。情報量の値を基に重要単語を選んだ場合と、出現回数や出現ページ数を基に重要単語を選んだ場合と比較すると、情報量の値を基に選んだ方が良好な結果が得られることが表 1 からわかる。

表 1: 重要単語の比較 (PRINTER)

順位	情報量	出現回数	出現ページ数
1	用紙	実現	実現
2	プリンタドライバ	対応	対応
3	解像度	BJ	採用
4	紙	印刷	最大
5	プリント	PC	用紙
6	実現	最大	解像度
7	印刷	プリンタ	こと
8	ハガキ	エプソン	搭載
9	カートリッジ	EPSON	印刷
10	用紙サイズ	こと	使用
11	OHP	プリント	プリンタ
12	給紙方式	MJ	税別
13	IEEE	採用	紙
14	プリンタ	オプション	機能
15	採用	使用	接続
16	ランニングコスト	PC-PR	オプション
17	印字	BC	幅
18	PC-PR	搭載	ため
19	印字速度	環境	プリント
20	インク	カートリッジ	環境

3 情報量を利用した再分類

3.1 特徴ベクトルの生成

2 章で求めた、それぞれの概念の構成要素を用いて、WWW ページの再分類を行った。その際に、それぞれの概念に関する単語の重みを要素として表現する概念ベクトルを生成し、それを利用した。特徴ベクトルには、各 WWW ページに対するものと、各概念に対するものの二種類がある。

まず、ベクトル空間モデル [1] を応用して、各 WWW ページの特徴ベクトルを生成する。 2 章で情報量を基に WWW ページから抽出した単語リストをキーワードとして、各キーワードが WWW ページに出現する頻度をその WWW ページの特徴ベクトルに対応する成分の値とする。こうしてできたベクトルを正規化したものを、その WWW ページの特徴ベクトルと定義する。ある WWW ページの特徴ベクトル $\boldsymbol{D}_n$ は次のように書ける。

$$\boldsymbol{D}_n = \frac{(term_{n11}, \cdots, term_{n1t}, \cdots, term_{ni1}, \cdots, term_{nit})}{\sqrt{\sum_{j=1}^{t}(term_{nj})^2}}$$

$term_{nij}$ = ある概念に属する WWWページnの$term_{ij}$の出現頻度

また、各概念の特徴ベクトルは、対応する各概念の構成要素を表す単語が出現する WWW ページの特徴ベクトルを平均したものとする。概念の特徴ベクトル $\boldsymbol{G}_i$ は、次のように書ける。

$$\boldsymbol{G}_i = \frac{\sum_{k=1}^{n}\boldsymbol{D}_k}{n}$$

3.2 特徴ベクトルを用いた再分類

2 章において、情報量の計算をするために、 手動で WWW ページの分類を行った。これらの WWW ページに対して、 3.1 節で求めた特徴ベクトルを用いて再分類を行い、情報量から求めた各概念の構成要素で WWW ページをどの程度分類できるのかを実験した。

WWW ページの特徴ベクトルと、概念の特徴ベクトルを用いて、該当する WWW ページと概念との類似度を求めた。類似度は、 WWW ページの特徴ベクトルと概念の特徴ベクトルとの内積で定義する。すなわち、 $\boldsymbol{D}_n$ という特徴ベクトルで表される WWW ページと、 $\boldsymbol{G}_i$ という特徴ベクトルで表される概念との類似度 $\mathrm{sim}(\mathrm{D_n}, \mathrm{G_i})$ は、

$$\mathrm{sim}(\mathrm{D_n}, \mathrm{G_i}) = \boldsymbol{D}_n \cdot \boldsymbol{G}_i$$

と表される。それぞれの WWW ページについて、各概念との類似度を計算し、最も類似性が高くなる概念に、その WWW ページを分類する。このようにして分類した結果を、適合率、再現率の二つの観点から評価した。なお、適合率、再現率はそれぞれ次のように定義される。

$$\text{適合率} = \frac{\text{正しく分類されたページ数}}{\text{分類されたページ数}} \cdot 100(\%)$$

$$\text{再現率} = \frac{\text{正しく分類されたページ数}}{\text{分類されるべきページ数}} \cdot 100(\%)$$

パソコン関連製品の情報に対して再分類を行った結果を表 2 に示す。

表 2: 同一データ使用時の適合率と再現率

カテゴリ	適合率	再現率
CD-ROM	89%	100%
PRINTER	96%	92%
SCANNER	100%	90%
WORD-PRO	98%	90%
MO	85%	94%
MODEM	93%	93%
TA	97%	92%

どの概念に対しても、適合率、再現率ともにかなり良好な結果が得られた。このことから、 WWW ページから抽出された情報量の多い単語を利用して、 WWW ページを分類することが可能であると言えるであろう。だが、この実験では、情報量の多い単語の抽出に用いたデータと再分類したデータが同一のものであるため、この結果は当然の結果とも考えられる。そこで、情報量の計算に用いたデータ (以下、 DataA) とは異なるデータ (以

表 3: 異なるデータ使用時の適合率と再現率

カテゴリ	適合率	再現率
CD-ROM	92%	94%
PRINTER	91%	78%
SCANNER	82%	96%
MO	66%	86%
MODEM	92%	66%

下、 DataB) に対して、同様の実験を行った。その結果を表 3 に示す。

この結果、概念によってはそれほど高い確率では分類されなかったものも見られたが、大体において高い確率で分類されていると言える。このことからも、本手法が情報の分類に対して有効であることがわかる。また、DataB に対して情報量の計算を行い、情報価値の高い単語を抽出したところ、 DataA から抽出された単語とは異なるものもいくつか出てきた。したがって、抽出される情報価値の高い単語を、分類するデータを別のものにした時にも利用価値が高いものにするためには、情報価値の高い単語を抽出する際に利用するデータの量を多くして、各概念に対して普遍的に存在する単語を抽出することが必要である。

4 概念獲得支援の実験

3.2 節の結果から、情報量を基に抽出した単語を利用した情報の分類が有効であることがわかった。では、本手法と、手作業で作成したオントロジーを利用して情報を分類する方法とでは、その分類精度にどのぐらいの差が出るのだろうか。手作業で作成したオントロジーを利用した情報分類の実験 (以下、手作業の実験) が既に行われている [5]。そこで、手作業の実験で用いられたデータと同一の、旅行に関する情報を用いて、情報量を基にした重要単語抽出を行った。また、抽出された単語を基に特徴ベクトルを生成し、それを利用して情報の再分類を行った。

表 4: 旅行に関する適合率と再現率 I

カテゴリ	DataC		DataD	
	適合率	再現率	適合率	再現率
行事	92%	67%	89%	83%
ホテル	88%	100%	63%	71%
神社	100%	86%	82%	64%
公園	87%	93%	60%	64%
交通	80%	94%	82%	82%
祭り	81%	100%	64%	82%
見所	83%	83%	68%	72%
店	100%	97%	97%	91%
温泉	88%	100%	87%	87%
施設	88%	100%	92%	79%
寺	95%	94%	93%	82%

情報価値の高い単語の抽出に用いたデータ (以下、 DataC) と同一のデータを再分類した結果と、 DataC とは異なるデータ (以下、 DataD) を分類した結果を表 4 に示す。なお、 DataC は手作業の実験に用いられたデータと同一のものである。

表 5: 手作業によるオントロジーを用いた実験結果

カテゴリ	適合率	再現率
宿泊・ホテル・旅館	100%	63%
交通	47%	66%
見所	100%	66%
行事	100%	70%
祭	100%	80%
公園	84%	81%
神社	96%	59%
寺	90%	65%
温泉	94%	76%
店・レストラン	100%	33%
施設	51%	71%

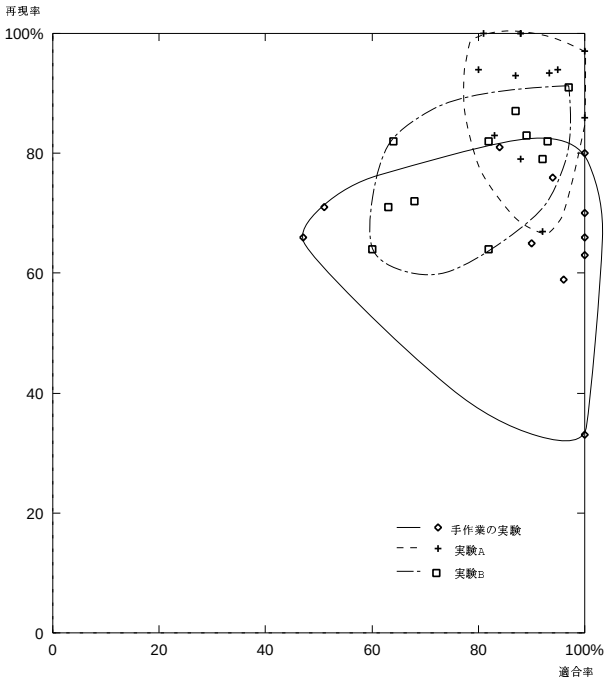

図 1: カテゴリの適合率と再現率の分布 I

これらの結果から、いくつかの概念で精度が低いものの、全体的には比較的精度のよい結果が得られている。このことから、本手法がある特定の分野にのみ有効なのではなく、様々な分野において有効であると言える。なお、“ホテル”、“公園”、“見所” といった概念の分類精度が低かった原因としては、概念と明確な実態との結びつきが弱いことや、他の概念との明確な境界が存在しないことなどが考えられる。

次に、この結果を手作業の実験の結果と比較した。なお、手作業の実験の結果は表 5 に示した通りである。適合率を縦軸に、再現率を横軸に取ってそれぞれの実験結果を示したものが図 1 である。なお、図 1 において、“実験 A” は DataC を再分類した結果を、“実験 B” は DataD を分類した結果をそれぞれ示している。

表 6: 旅行に関する適合率と再現率 II

カテゴリ	DataC		DataD	
	適合率	再現率	適合率	再現率
行事	90%	50%	83%	56%
ホテル	88%	88%	57%	50%
神社	85%	79%	83%	71%
公園	72%	93%	79%	79%
交通	73%	94%	86%	71%
祭り	82%	82%	82%	82%
見所	71%	56%	50%	50%
店	100%	97%	93%	76%
温泉	81%	87%	76%	87%
施設	86%	34%	95%	72%
寺	91%	97%	84%	66%

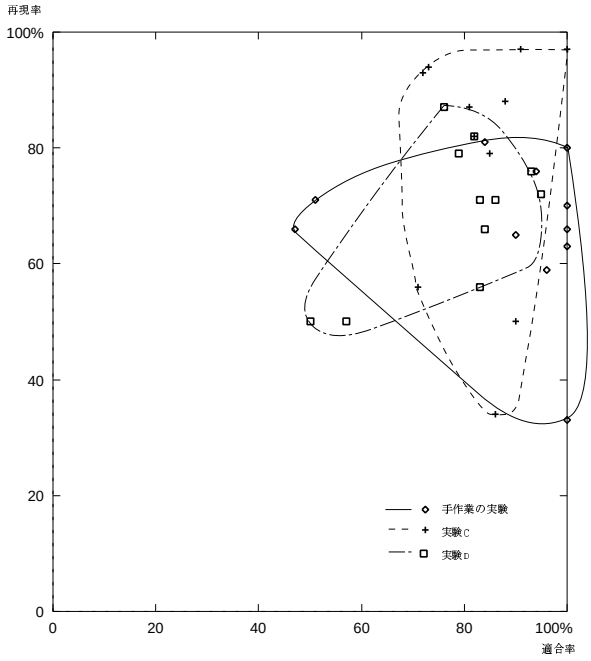

図 2: カテゴリの適合率と再現率の分布 II

この結果から、本手法を用いた結果が手作業の実験結果よりも高い精度で分類されている、と言える。しかし、実験 A 、 B で用いた特徴ベクトルの成分数が 220 個であるのに対して、手作業の実験で用いられた特徴ベクトルの成分数は約 50 個であるため、この結果だけからでは、本手法のほうが優れている、とは言い切れない。そこで、特徴ベクトルの成分を 55 個とし、再度同様の実験を行った。その結果を表 6 に示す。また、適合率を縦軸に、再現率を横軸に取ってそれぞれの実験結果を示したものが図 2 である。なお、図 2 において、“実験 C” は DataC を再分類した結果を、“実験 D” は DataD を分類した結果をそれぞれ示している。

この結果を見ると、本手法と、手作業の実験では、ほぼ同程度の精度で分類されていることがわかる。このことから、手作業の実験では手動で作成したオントロジーを用いていることを考慮に入れると、情報量を用いて重要単語の抽出を行う方法がとても有効であることがわかる。

5 おわりに

ある概念に関連のある単語を情報量を用いて抽出する手法を提案した。 WWW 上の情報から本手法を用いて抽出された単語を利用して情報の分類を行った結果、高い精度で分類されており、本手法が概念に関連する単語の抽出に有効であることがわかった。また、このことから本手法により得られた単語は、概念獲得支援になる。

現在のシステムは、類義語には対応できていない。ある二つの単語が同じ事柄を指していても表現が異なれば別の単語とみなされてしまうため、出現頻度が落ち、重要単語から漏れてしまっている可能性がある。同一の事柄を指す単語群を同一の単語とみなせるようにすれば、抽出された単語の価値がさらに高まるものと考えられる。

また、一つの WWW ページは必ず唯一つの概念に分類されるため、旅行に関する概念のような概念間の明確な境界線がない場合は、分類精度が低くなってしまう。概念間の明確な境界がないような分野に対しても高い分類精度を保つためには、一つの WWW ページをある一定以上の類似性を持つ複数の概念に分類する、といった方法を用いる必要があると考えられる。

参考文献

[1] G.Salton and M.J.McGill. Introduction to Modern Information Retrieval. *McGraw-Hill*, 1983.

[2] Michael Pazzani, Jack Muramatsu and Daniel Billsus. Syskill & Webert: Identifying web sites. *AAAI-96/IAAI-96 Proceedings Volume One*, pp.54-61, 1996.

[3] 岩爪 道昭, 白神 謙吾, 畑谷 和右, 武田 英明, 西田 豊明. オントロジーに基づく広域ネットワークからの情報収集・分類・統合化. 情報処理学会論文誌, 1997.

[4] 大杉 英一. ネットワークからの情報抽出によるオントロジー獲得支援. 奈良先端科学技術大学院大学 情報科学研究科, 修士論文, 1997.

[5] 白神 謙吾. インターネットにおける情報分類とオントロジー獲得. 奈良先端科学技術大学院大学 情報科学研究科, 修士論文, 1996.