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

2004－DPS－118 (13)
2004／6／4

電荷式を用いた情報選択システムの提案

斉藤 研一郎* 井前 吾郎*
重野 寛* 岡田 謙一*
{saitoken, inomae, shigeno, okada}@mos.ics.keio.ac.jp

近年インターネットの発展などによって、我々の身の回りには大量の情報が溢れている．しかし、あまりに情報が多いために自分にとって本当に有益な情報を見落としてしまう可能性が増えている．そこでgoogleなどのように情報を検索、選択するシステムが必要不可欠になっている．そして近年では、単純に情報を選択するシステムだけでなく、ユーザの興味を自動的に解析し、情報を自動的に選択するシステムに注目が集まっている．そこで本稿では、ユーザの興味を正確に、そしてリアルタイムに解析し、ユーザにとって有益な情報だけを選択することを目的とした、電荷式を用いた情報選択システムを提案する．

Information Selection by Electric System

Kenichiro Saito* , Inomae Goro* ,
Hiroshi Shigeno* , and Ken-ichi Okada*

Recently, depending on the Internet development, there is a ton of infomation in our environment. But we can't find infomation of great benefit for too much information. And so we need the system that retrievals and selects the information. Especially, it is not enough to the simple selection system, the system that analyzes an user's interest and selects the information automatically. This paper discribes the information selection system that uses the electric charge to analyze the user's interest precisely and in real time, and selects only the information of gerat benefit for an user.

1 はじめに

近年インターネットの発展によって、我々の身の回りには大量の情報が溢れている．そして、googleなどの検索システムによって、それらの情報を得たいときにいつでも得られる環境は整いつつある．しかし、情報があまりに多すぎるため、検索をしても関係のないたくさんの情報が選択され、ユーザは自分にとっての必要な情報が何かわからかくなり、結局必要な情報を得ることができないことが多くなってしまっている．そのため、多くのユーザが自分自身にとって有益な情報を見落とす可能性が増えてしまっている．そこで、そういった問題点を解決するためにユーザの興味を発見し、それに従い有益な情報だけを選択するシステムに注目が集まっている．

そして今日では、インターネット放送やデジタル放送でのユーザ興味にあった広告の選択や番組選択システムに代表されるように、過去の履歴情報を用いた情報選択手法の研究が進められている．つまり、ユーザのホームページや検索状況等の履歴情報やＴＶ番組の視聴履歴を解析しユーザの興味を発見し、それに基づいてユーザの興味がある情報を選択しようというシステムである．しかし、それら既存の多くの方法が単純なキーワード検索を行っている場合が多く、それでは正確にはユーザの興味を発見できない．さらに大量の履歴情報が必要になるためリアルタイムな情報選択が行えないという問題点がある．

そこで我々は、それらの問題点を解決することを目的に、ユーザの興味を正確に、しかもリアルタイムに発見する手法として電荷式を用いた情報選択手法を提案する．そして、現在のインターネット状況の縮図と

* 慶應義塾大学大学院理工学研究科開放環境科学専攻
School of Science for OPEN and Environmental Systems,
Graduate School of Science and Technology,
Keio University

して受験生の大学選択を考え、大学紹介システムを実装し、本手法の評価を行った.

以下, 2章で提案する電荷式を用いた情報選択手法の特徴, 3章では実装したシステムの概念と流れ、4章ではシステムの評価について述べ、5章を結びとする.

2 電荷式の特徴

上述したように、履歴情報を用いた情報選択手法[3],[4],[5]は様々な研究がなされているが、それらの多くは以下のような問題点がある.

- ユーザがあらかじめ入力しておいた年齢・性別などのデータと履歴情報を比較して情報選択をするので、非常に面倒であり、さらに初めてのユーザは使用することができない
- 履歴情報にこだわり過ぎる余り、過去の特性に縛られ、ユーザの興味に動的に対応することができない
- 過去の特性による興味発見では、情報配信者側にとっても、新しい情報やあまり有名でない情報を提示しにくくなってしまう

そこで、我々は上記のような問題点を解決する手法として電荷式を用いることを考えた. そこでまず電荷式について説明する. 電荷式とは、一般的にクーロンの法則[1]と呼ばれており、二つの電荷 q_1, q_2 に働く力を示している. そして、クーロンの法則は以下の式で表される.

$$F = \frac{1}{4\pi\epsilon}\frac{q_1 q_2}{r^2}$$

そして, 回りに並んだ4つの点電荷 Q_1, Q_2, Q_3, Q_4 があり, それらは固定されているとする. 電荷量はそれぞれ q_1, q_2, q_3, q_4 とする. この中に, 電荷量 $-q$ の点電荷qを落とすと, qに働く力は, 下記のようになる.

$$F = \frac{1}{4\pi\epsilon}\frac{q_1(-q)}{r_1^2}+\frac{1}{4\pi\epsilon}\frac{q_2(-q)}{r_2^2}+\frac{1}{4\pi\epsilon}\frac{q_3(-q)}{r_3^2}+\frac{1}{4\pi\epsilon}\frac{q_4(-q)}{r_4^2}$$

ただし, Q_1, Q_2, Q_3, Q_4 とqの距離をそれぞれ r_1, r_2, r_3, r_4 としている.

ここで上記の式より、2次元のx y座標軸上でrを計算すれば, F = 0となる点, つまりqが Q_1, Q_2, Q_3, Q_4 のどの点電荷からも釣り合っている点は一意に決めることができることがわかる (図1).

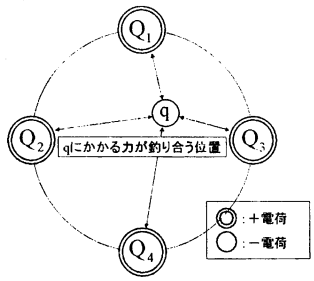

図 1: 一意に決まる電荷

つまり、電荷式を用いれば多次元の特性を持ったものも、二次元で表すことができると考えられる. さらに電荷は、自由に2次元上を動くことができるので、動的にユーザの興味を発見するのに適していると考える. そういった性質を活かし、本手法は以下のような特徴を持っている.

- あらかじめユーザの入力がなくても、履歴情報から電荷量を変えるだけですぐにユーザの興味を解析することができる
- 過去の特性に捉われず、新たな特性が加わった場合にも、動的に対応することができる
- 新しい情報もすぐに付け加えることができるので、情報配信者側にとっても、ユーザに対して容易に自分達の情報を伝えることができる

3 システムの実装

3.1 概念

上述した電荷式の特徴を活かし、本研究では大学紹介システムの実装を行った. なぜ大学紹介システムにしたかというと、一般的な受験生にとって、大学の数は非常に多く、その差異がわかりにくい. つまり、実際自分に合っている大学がどこなのか、どこからの情報を得ればいいのかがわかりにくいという現実がある. そういった状況を、大量の情報に溢れた現在の情報化社会の縮図と捉え、システムの実装を行った. 大学は、関東にある大学のデータ (102大学) を使用した[2].

具体的に、上述の電荷式の特徴を大学紹介システムに適用すると以下のようになる.

- ユーザが面倒な入力をすることなく、自分の興味ある大学のホームページを閲覧しているだけで、システムはユーザ興味を発見することができる

- ユーザの興味ある大学が変わったとしても、その場に応じて電荷の値を増減させるだけで、動的に興味を解析することができる
- 新らしい大学ができた時、新たにその大学の電荷を作成するだけで簡単に新たな情報を加えることができる

3.2 システムの流れ

それでは、実装したシステムの流れについて説明する.

1. 各大学は、偏差値、人数、所在地などといった様々な特性が違うと考えられる. そこで、本システムは表 1 のように各大学 (ここでは a〜g 大学) において、例えば偏差値、人数といった様々な特性 (ここでは特性 α〜ζ) に対するデータを表として作成し、所持している.

表 1: 各特性に対する情報のデータ

情報名	特性α	特性β	特性γ	特性δ	特性ε	特性ζ
a	5	0	5	0	1	2
b	5	0	0	5	4	4
c	5	0	0	5	3	3
d	0	5	5	0	2	5
e	0	5	0	5	1	2
f	0	5	0	5	5	2
g	0	5	5	0	4	1

2. 図 2 のように、システムは各特性 α〜ζ (初期値 10) を円形に並べ (特性電荷)、その中に各大学の電荷 (情報電荷) を落とすという作業を行う. この場合、概念的には特性電荷をプラス電荷、情報電荷をマイナス電荷と考える. すると、各情報電荷は表 1 の値によって、それぞれ電荷の規則に従い、特性電荷が作る円内に一意に位置を決める. それを、電荷の初期配置と呼ぶことにする.

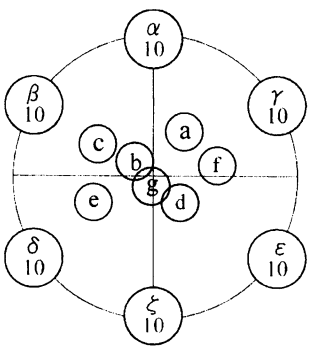

図 2: 特性電荷と情報電荷の初期配置

3. ユーザは、自由に各大学のホームページを閲覧する. すると、システムは閲覧時間順に大学のホームページを作成し、上位 3 大学をユーザが興味を持っている大学として抽出する. ここでは、g・d・a 大学が上位 3 大学であった、つまりユーザが興味を持った大学であるとする.

4. システムは表 1 のデータから g・d・a 大学の値を抽出し、初期値が 10 であった特性電荷にそれぞれの値を足すという作業を行う. 例えば、特性 α は g 大学:5,d 大学: 0,a 大学:0 であるから、10(初期値)+5+0+0=15 となる. それを α〜ζ まで全ての特性について行い、そして値の大きくなった特性電荷を大きい順に上から並べ替えるという作業を行う. この場合、特性 γ が 10+5+5+5=25 と一番大きいので一番上に来ている. なぜ並び替えを行うかというと図 3 のように、極端の特性が大きくなってしまった時に、情報電荷の位置がちょうど真中になり、どの特性の影響を受けているのかわからなくなることを避けるためである. すると、特性電荷の配置がユーザの興味を考慮した配置に変わる (図 4).

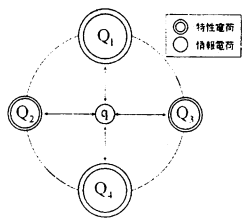

図 3: 並び替えがない場合の問題

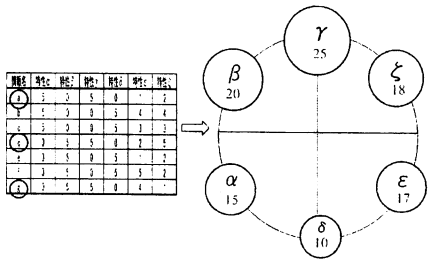

図 4: 特性電荷の変化

5. 2と同じように、再び表1の値を持った各大学の電荷を落とすと、先ほどの初期は配置とは異なったユーザの興味を考慮した大学電荷の配置が決定する(図5). そして、ユーザが現在a大学を閲覧しているとすると、図5においてa大学に距離的に近い大学を紹介する. ここで「距離的に近い大学」というのは、このユーザにとって似ていると感じると思われる大学を表している. 例えば、現在a大学を閲覧している場合、図5よりbとf大学が距離的に近いので、このユーザにとってはa大学とb・f大学は似ていると感じるだろうとシステム側は判断するということになる. つまり、このユーザのみの完成において、ある大学に似ている大学を紹介するということになる.

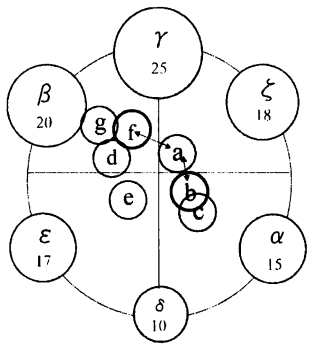

図5: ユーザの特性を考慮した配置

3.3 実装画面と操作

本システムの実装画面は, 図6のようになっている.

1. ユーザはまず、左のフレーム(1)から自分の興味がある大学を選択し、そのホームページを閲覧する. するとシステム側では、ホームページの閲覧時間に応じて、ユーザが興味あると考えられる順番に大学のランキング付けを行う.
2. ランキングの上位3大学を抽出し、電荷式のシステムに従って、ユーザ独自の電荷マッピングを作成する.
3. 電荷マッピングに従い、現在ホームページを閲覧している大学に近い5校の大学とユーザの特性(2)を表示する.
4. ポップアップ画面(3)には、下記に記述する12項目に対して、ユーザがどれだけ興味があるかを棒グラフで示している.

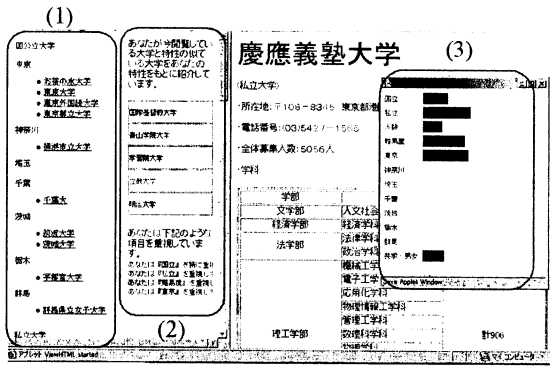

図6: システムの実装画面

4 システムの評価

4.1 大学データ

実際システムを実装する際に、特性電荷として以下の項目を使用した.

- 国立(公立)か、私立か(2項目)
- 人数
- 難易度
- 所在地(東京、神奈川、埼玉、千葉、茨城、栃木、群馬:7項目)
- 共学か、その他

それぞれの項目は、重要度が違うと考えられる. そこで、どの項目を重要視するかを事前アンケートによって求め、そしてその結果をもとに、適当な値を算出し各特性について正規化を行った. 順番及び正規化の基準値は以下のようになった.

1. 難易度(偏差値のデータをもとに30に正規化)
2. 国立(公立)か、私立か(それぞれを15とする)
3. 人数(200人単位で1とする)
4. 所在地(それぞれの県にあることを5とする)
5. 共学か、その他(それぞれを10とする)

そして、各大学についてそれぞれの項目の値を算出し、データベースとしてシステムは所持している. 大学データの例を表3に示す.

表 2: 大学データ

大学名	大学コード	国立	私立	人数	難易度	東京	神奈川
東京	1	15	0	8.135	30	5	0
東京外語	3	15	0	1.8625	22.5	5	0
東京都立	19	15	0	1.875	23.4375	5	0
国際基督教	45	0	15	1.55	30	5	0
慶應	41	0	15	12.64	30	5	0
早稲田	129	0	15	22.1	25.5	5	0

そして今回の評価では、大学のホームページの情報量の違いによって、ユーザに対する影響が異なるのを防ぐため、各大学について同じ情報量を持ったホームページを作成した．ホームページには、上記で示した特性電荷の項目についての情報の他に、受験科目などユーザが受験するために必要な情報が書かれている．そのホームページをもとに、ユーザに興味ある大学を閲覧してもらう．

4.2 評価方法

システムの有用性を示すため、以下の 3 つの評価を研究室の学生 20 人に対して行った．

1. 電荷式の有用性、及び初めてのユーザに対する正確な情報選択に対する評価
   あらかじめ指定した大学を選択した場合、紹介された 5 つの大学のうち、いくつ適当であると感じるかをアンケートにより述べてもらう．

2. ユーザの興味にあった情報を提供できたか、また動的にユーザの興味を解析できたかということについての評価
   実際ユーザに自由に興味がある大学のホームページを 4 つ以上閲覧してもらい、そこで紹介される 5 つの大学の内、いくつ似ていると思ったかについてアンケートにより述べてもらう．

3. 現実に受験生が使用している方法との比較アンケート評価

   ここでいう「現実に受験生が使用している方法」とは、今回データを入力するために使用したような受験本によって偏差値、場所などにより自分の興味がある大学を探す．そして、その大学に似ている大学をインターネットを用いて、単純なキーワード検索を行い探すというものである．その既存手法と、本システムを両方しようしてもらい、アンケート評価を行った．アンケート項目は下記に記す．

4.3 評価結果及び考察

4.3.1 電荷式に関する評価からの考察

上記 1. の評価を行った所、紹介された 5 つの大学の内 3.9 大学、つまり 78 %の大学を似ていると判断することがわかった．このことから、電荷式を用いたこのシステムでは、初めてのユーザであっても、ほぼ 80 %の確率でに興味を捉えることができることがわかった．

4.3.2 動的なユーザ興味解析に関する評価からの考察

上記 2. の評価を行った所、紹介された 5 つの大学の内 3.5 大学、つまり 70 %の大学を似ていると判断することがわかった．評価 1. でも、大学によって似ていると思う大学にばらつきが見られたが、この評価ではより顕著にばらつきが見られた．ユーザの閲覧状況により、より正確に興味を発見できると考えたが、評価 1. よりも値が下がってしまったのは残念であった．ランキングの抽出順位を 3 位までではなく，5 位までなど多くしたらより正確な結果が出ると考える．

4.3.3 従来手法との比較アンケートの結果からの考察

上記 3. のアンケート評価に対する項目と結果は以下の表 3 のようになった．

表 3: アンケート結果

	既存システム	本システム
①次に閲覧する大学を見つけやすい	1.7	4.6
②自分が何に興味があるかわかりやすい	1.8	4.2
③インターフェースが見やすい	2.2	4.3
④大学のHPはわかりやすい	2.8	4.0
⑤あまり知らない大学を発見できる	1.2	3.1
⑥本手法が現在のネット状況に使いたい	2.0	4.2
⑦興味の変化に対応している	-	4.3
⑧提案された大学を受けてみようと思った	-	4.2
⑨紹介された5大学は興味にあっている	-	4.6
⑩興味を表す棒グラフは有効である	-	4.6

評価項目 1・8・9 より，本システムの方が興味を発見し，情報を選択しているため，従来の方法よりも情報検索が容易になっていることがわかった．また 7 により，上記で示した動的なユーザ興味解析は，アンケートからも実証された．そして 2・10 より，ユーザの興味発見にも本システムは有用であることがわかった．2 章で示した問題点も、既存方法より大きな改善を見ることができた．

また 5 より，あまり数値は高くはないが，あまり有名でない大学をユーザ側に知らせることもできることがわかった．つまり本システムは，ユーザ側だけでなく配信者側にとっても，新しい情報，または無名な情報を伝えるのに有用であることが示された．

5 結論

5.1 まとめ

インターネットの発達によって，身の回りに大量の情報が溢れるようになった結果，自分にとって本当に必要な情報を見落とす可能性が増大した．そこで近年は，マイニングの技術を活用し，ユーザの興味にあった情報だけを選択するというシステムに注目が集まっている．しかし，既存の手法では以下のような問題点がある．

- 煩雑なデータを入力する必要があるものが多く，初めてのユーザには使用しづらい
- 動的に変わるユーザの興味に対応できない
- 配信者側にとって，新しい情報や無名な情報をユーザに提示することが困難である

そこで，我々は，既存手法の問題点に考慮し，電荷式を用いた情報選択システムを提案した．そして，現在のインターネット状況の縮図として，大学紹介システムを実装し評価を行なったところ，上記の問題点を大部分改善されたシステムであることがわかった．

以上より結論として，本研究はデータをきちんと集めれば，現在の複雑で大規模なインターネット状況であっても，ユーザ興味を発見することができ，有用なシステムになり得ると考えることができる．

5.2 今後の課題

上記のような有用性が考えられる一方、今後の課題も多く考えられる．例えば，データの正確さ、項目をどのように決定するかなど様々上げられる．

そしてその中でも、もっとも考えなければいけないユーザ興味は定量的なものだけでは推し量れないということである．今回実装した大学紹介システムでも，前述した１２個の項目だけでなく、大学には雰囲気というものがあり、それはユーザの興味に大きく影響するものである．しかし、それをシステムが認識するのは非常に難しい．そのような定性的なデータを扱えるようにならなければ，本当に正確にユーザ興味を解析することはできないと考える．今後は、定性的なデータを扱えるようなシステムを考えていかなければいけないと考える．

また、システムに対するきちんとした評価方法に関してもこれから考えていかなければいけないと考えられる．ユーザの興味の発見など定性的な評価に関して現状では、システムを実際使ってもらいアンケートによる評価しかない．それではきちんとした評価が取れているか非常に曖昧であると考えられるので、今後は定性的な事柄に対する評価も考えていかなければならない．

参考文献

[1] Kimihisa Sakima.,Coulombu'sLow., http://www12.plala.or.jp/ksp/formula/physFormula/html/node20.html,2003

[2] 学習研究社，２００４年度用　大学の学科比較案内,2003

[3] 山田和弘, 片山修一, 寺西裕一, 奥田剛, 下条真司, 宮原秀夫. コンテンツ配信におけるメタデータに基づいたＣＭ選択機構の提案と評価,DICOMO,2003

[4] 中島太郎, 渡辺尚, 樽口秀昭. 履歴情報を用いたＴＶ番組選択支援エージェント, 情報処理学会論文誌,Vol.42,No.12,December

[5] 織田充, 南俊朗, 有馬淳. 検索ログを用いたキーワード推薦エージェント, 情報処理学会研究報告