

## 意外な検索キーワードを推薦する手法の提案

中村 将人<sup>†</sup> 木村 昌臣<sup>†</sup>

芝浦工業大学工学部情報工学科<sup>†</sup>

### 1. 背景と目的

ユーザが自分の興味のある事柄を調べるために、検索サイトが利用されることが多い。その際にユーザは自分の知っているキーワードを入力して検索を行うが、自分の知らないキーワードについての事を調べるのは困難である。

先行研究[1]では全ユーザの検索履歴から、キーワード同士をエッジで結びネットワークを構築し、ユーザが検索したキーワードと直接エッジで結ばれたものの中から意外であると考えられるキーワードを取得したが、実際にはすでにユーザが知っているキーワードを推薦してしまうことが多かった。そこで、間接的な関連からキーワード群を抽出することでより意外であると思われるキーワードを抽出できると考えた。

本研究では「意外性を持った検索キーワード」を「調べたい事柄と弱い関連があるもの」と定義する。提示されたものが調べたい事柄と全くの無関係ならば検索ワードとして適していないと考えられ、逆に関係が強いものは、ユーザが予測できてしまう。以上をもとに「意外性を持った検索キーワード」の推薦手法の提案を行う。

### 2. 提案手法

複数のユーザの検索履歴からネットワークを構築し、ユーザが調べたいキーワードと弱い関連のある意外性を持ったキーワードを取り出し推薦する。

#### 2.1 検索キーワードのネットワーク構築

全ユーザの検索履歴から、同時に検索されたキーワードは関連を持っているとする。キーワードをノード、キーワード同士の関連をエッジとしたネットワークを構築する。例えば「芝浦工業大学」と「豊洲」が同時に検索されたログがあった場合、その2つをノードとして、間にエッジを張ることでネットワークが構築される。

#### 2.2 間接的な関連を持つキーワードの取得

ネットワーク上で、ユーザが検索語として使いたいと考えているキーワードのノード(以後、興

味キーワード)と直接つながっていないが1つのノードを経由してつながっているノードを間接的な関連を持つキーワードとして取得する。

#### 2.3 意外性を持ったキーワードの抽出

興味キーワードから間接的な関連を持つキーワードを取得すると、その数は膨大なものになる。また2つ離れているというだけでは、興味キーワードを言い換えただけのキーワードが得られてしまう恐れがある。そこで抽出されたキーワード群と、ユーザが調べたいキーワードの間にどの程度の結びつきの強さがあるのかを以下の指標を用いて求める。

$$\text{意外性の強さ} = \left( \sum_{i=1}^N M_i \right) \frac{K}{N}$$

$N$  は、意外性の強さの値を調べたいノードと、ユーザの興味キーワードのノードの双方とエッジで繋がっているノード(中間ノード)の数である。 $N$  の値が大きいと、それだけユーザの興味キーワードと意外性の値を調べたいキーワードの間に共通点が多くなる。また、ユーザの興味キーワードと同じ意味を持つ言葉を言い換えたものであるという可能性も高くなり、それだけ意外性が弱くなると考えられる。

$M_i$  は各中間ノードが持つ次数(結びついているエッジの数)である。この次数が大きいと、ユーザの興味キーワードと意外性の値を調べたいキーワードの間にあるノードが様々なノードからエッジをつなげられていることになる。そのため、間にあるノードは様々な事柄と関係があることになり、ユーザの興味キーワードと意外性の値を求めたいノードの2つを意外な面から間接的に結びつけることが期待できる。

$K$  は、意外性の値を調べたいノード自体が持っている次数である。 $K$  の値が大きい場合にはノード自体が様々なノードと関連しており、 $K$  の値が小さいノードと比較して興味キーワードのみとの関連が相対的に弱くなる。関連が弱くなることで、意外性を持った検索キーワードとなることが期待できる。

本手法に基づく意外性の強さの値の大小関係についての模式的な説明を図1に示す。

Proposal to recommendation technique with surprisingness.

<sup>†</sup>Masato Nakamura, <sup>†</sup>Masaomi Kimura

<sup>†</sup>Shibaura Institute of Technology

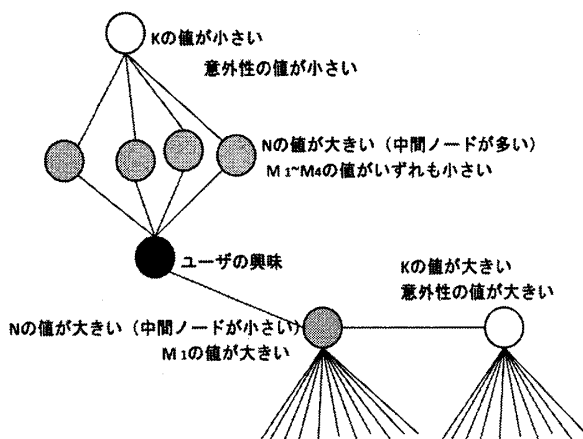


図 1 意外性の値の大小

### 3. 実験

#### 3.1 実験対象・概要

本研究の実験では「芝浦工業大学」に直接的に又は間接的に関連のあるキーワードに限定して実験を行った。このような制約を設けるのは、別々の分野で検索が行われた場合にノードが孤立しネットワークを構築できない状況になるのを避けるためである。

実験は 2 段階に分けて行った。最初に同時に検索されたキーワード同士を結ぶネットワークの構築を行い。その後、構築されたネットワークから意外性を持った検索キーワードを抽出できているかのアンケートすることで評価を行った。

ネットワークの構築については、本学学生 6 人に協力を依頼し「芝浦工業大学」に関連がある複数のキーワードを用いて検索を行ってもらった。これによって「芝浦工業大学」を中心とするネットワークを構築した。

続いて、入力したキーワードから本手法を用いて得られる意外性を持つキーワード群をもとに仮想的な「入力キーワード」との関係を開く 2 択の問題を 20 組作成した。各組に対して入力キーワードから「どちらが容易に想像できたか」「どちらがより関連性があるか」「どちらがより意外であるか」の 3 つの質問を行った。問題は 2 パターンの問題を 10 組ずつ用意した。すなわち、入力キーワードに対して出力された意外性を持つキーワードの中から意外性の値が高いものと低いものを 1 つずつ選んだ質問(タイプ 1)を 10 組、本手法で得られた意外性を持つキーワードから 1 つ、もう片方を中間ノードから選んだものを 1 つずつ選んだ質問(タイプ 2)を 10 組とした計 20 組である。問題はこの 20 組をランダムに並び変えたものを 13 名の被験者に提示した。

#### 3.2 結果・考察

表 1 は、20 組の質問に対して本研究の手法にお

いて意外性を持つキーワードであると提示されたキーワードが選ばれた割合である。「意外性の大小」は質問タイプ 1 において意外性の値が高いキーワードが選ばれた割合であり「隣接ノード」は質問タイプ 2 において意外性を持ったキーワードとして出力したものが選ばれた割合である。

表 1 アンケートによる回答の割合

(%)	想定できない	関連ある	意外だ
意外性の大小	52.0	44.7	53.7
隣接ノード	72.9	24.8	73.6

意外性の強さの値が大きいキーワードは、値が小さいキーワードに比べて「想定できるキーワード」や「関連性が高いキーワード」としては比較的选择されにくく「意外性があるかどうか」については約 54%が選択されたことから、指標の大きさに応じた意外性を感じている傾向がある程度見られたと考えることができる。隣接ノードに当たるキーワードと比較した場合には意外なキーワードであると回答される場合が多く、出てくるのが想定される可能性や、関連性があるとされる可能性が低いという結果が出た。

また、「想定ができ、関連があり、意外ではないキーワード」あるいは「想定できず、関連が薄く、意外と思われるキーワード」という回答が全体で 95.2%と非常に高い値を占めた。更に 13 人の被験者のうち 6 人が 20 組の質問全てに対して上記に当てはまる回答をしており、このことから「調べたい事柄と関連が弱いものは「推薦されてくるのが想定できないもの」であり「意外性を持ったキーワード」と考える人が多いという可能性が高いと考えられる。

### 4. まとめ

本研究の提案手法によって、意外性を持つキーワードを推薦するには、興味キーワードの隣のノードより、間接的に関連があるキーワードから推薦したほうが良い結果が出た。意外性の強さの指標については意外なキーワードの出る傾向は出たものの、改善の余地は大いにあると思われる。

#### 参考文献

- [1] 大塚浩平:意外性に着目した検索キーワードの推薦手法の提案,平成 20 年度芝浦工業大学工学部 情報工学科卒業論文.
- [2] 前田実香:言葉の関連性を用いたなぞかけ生成とその評価,筑波大学大学院博士課程システム情報工学研究科修士論文 (2005).
- [3] 村上知子,森紘一郎,折原良平:推薦結果の意外性を評価する指標の提案,人工知能学会第 21 回全国大会論文集, 2C5-2 (2007).