関心度予測によるWWW検索結果の絞り込み法に関する考察

2Y－10

山縣　修†　　津田 和彦†
筑波大学大学院 経営システム科学専攻

1.はじめに

最近、HTMLテキストをコンピュータネットワーク環境に構築したWWW(World－Wide Web)はGUIによるインタラクティブな操作により、様々な情報を手軽に得られるようになった。

しかし、HTMLテキストの基本は、リンクによりテキスト間を関連付けるハイパーテキスト形式である。ハイパーテキストは文書構造が非線形であるため、「迷子問題(どの場所にいるか、どう辿るか)」、「認知負荷問題(ラベルで次のリンクの価値を判断する)」が指摘されいるが、この問題の解決は充分に成されてはいない。

本稿では、WWWのドキュメントを対象に「利用者が情報を充分得たか、今後得られるか」の判断を予測する方法の一考察を与える。

更に、家庭電気製造業のサイトの一部製品に適用し、判断(関心度)を予測するモデルに指標が有効であることを検証した。

2.研究の背景

WWWが様々な分野に普及し、多くの利用者が拾い読みに相当する「ブラウズ」により情報を求めている。

しかし、ハイパーテキストの非線形構造(リンク構造)により、ある情報に到達するための経路が複数存在し、その経路を自由に選択できると同時に一つの遷移先しか認識できない。そのため、WWWはリンク構造を遷移している間に自分の位置を見失ってしまう「迷子問題」やリンクの先にあるノードの重要度をラベルだけで理解しなければならない「認知負荷問題」を継承している。

また、情報の視覚化の視点から大規模データベースを対象にした3次元視覚化の研究が進み、閲覧履歴の順序表示、ノード・リンク構造の視覚化等が提案されているが、新たに「表示速度の低下」、「認知負荷の増加」が明らかになり、「いかに多くの情報を表示するかではなく、いかに必要な情報だけを表示するか(表示情報量の削減)」が重要な課題と指摘されている。

一方、具体的な研究例である「納豆ビュー」[4]は、現時点では構造(ノードとリンク数)とノードに付属する属性の「持ち上げ操作」を基本として、利用者に優しい表示機能を実現している。しかし、この研究で提案された手法は簡便かつ有効に「迷子問題」を解決しているが、個々のノードが持つ情報を考慮しておらず「認知負荷問題」の解決に充分ではない。

3.研究の内容

WWWページは、ハイパーリンクにより関連情報へリンクされているため、そのリンクが無限に続く可能性がある。利用者は関心のある情報を得るためにWWWページを閲覧するのが一般的であるが、「感心ある情報が充分得られたかあるいは今後得られるか」は経験則により判断している。そのため、このサイトのブラウズを継続(ラベルをクリック)するか離れるか(他のサイトをブラウズするか、ブラウズを止めるか)の判断を誤り、充分な情報を得られない場合や、無駄に多くのページを閲覧する事による時間のロスを招いている。

本研究では、WWW閲覧時の無駄を無くするため、リンク先の単語の出現頻度とリンク数より今後得られる情報量を予測し、この結果より閲覧継続の指標を示すことを考察する。さらに、この指標により圧縮した情報を視覚化することで判断の負荷を軽減する可能性を家庭電気製造業のサイトの一部製品に適用し検証した。

3.1 指標

発想法における発散的思考と収束的思考、ノーランの概念モデルにおける実行の行為と評価の行為で示されるように、2つの過程で構成されることが多い。これらを発散過程(↗)と収束過程(↘)に対比させ、「上に凸」な形状と見做す。

A consideration about the focusing WWW-pages by using the degree prediction of concern.
Osamu Yamagata
Graduate School of Systems Management The University of Tsukuba.Tokyo

また、情報の広がりは同一レベルのリンク数、情報の深さはリンクのネスティングの深さに対応する。

一方、自然言語理解の研究において「ある文書の主題とその文書に現れる内容語(名詞、動詞他)には強い関連がある」ということが判っている。

そこで、ノードの情報とリンク構造を同時に数量化したノード単位の指標は、次の関係式で表わされる。

即ち、今後得られる情報量とは,関心のある事象に関する単語数と、リンク数の双方に相関する。よって、指標は下式により得られる。

指標＝単語群の出現頻度×リンク数

・単語群の出現頻度:利用者の関心を規定する単語(含む複数)の該当ノードでの出現頻度(の総和)。

・リンク数:関心に対応する情報の存在を明示する該当ノードでのリンクの数。(Top, Nextへのリンクは除く。)

更に、サイトの指標はブラウズしたノードの順序性を保ち、ノードの指標の列として表わす。

Site指標＝{Node1指標、……、Noden指標}

3.2 検証

日本の代表的製品である家庭用電気機器メーカーのサイトを対象に、製品の価格帯が近く、技術要素の全く異なる'成熟製品の冷蔵庫'と'今後の製品であるワイドTV'に適用した結果を図1, 2に示す。図から明らかなように、双方とも凸形状を示す2次元カーブヒッティングした形状を示した。

※ x-軸:閲覧したサイトの順序
y-軸:リンクの深さ(閲覧したノード順序)
z-軸:サイトの指標

※ 単語群は冷蔵庫をイメージする18語、ワイドTVの技術、規格(定格)と価格を現す11語

3.3 考察

各サイトのノードの単語群出現頻度とリンクの数を乗算した指標は、11サイトの内10サイトの指標は「上に凸」な形状を示した。

これにより、今後得られる情報量が当初は徐々に増加傾向にあり、ある程度ページ遷移をすすめると、減少することがわかる。よって、グラフの「上に凸」の頂上まで遷移を進めれば、ある程度の情報量が得られるとの指標を与えることが可能に為る。

尚、各サイトのリンクの深さは製品分類、リンクの広がりは製品数に対応している。また、下に凸形状を示した1例は、最深のノードで機能詳細を展開後に再度製品説明にリンクする構造である。

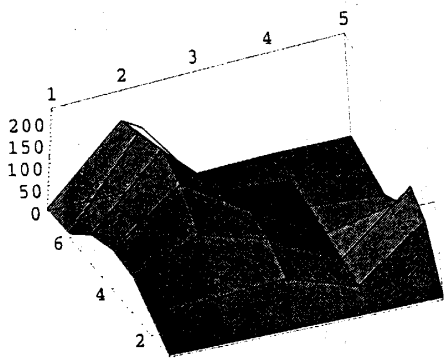

図1 冷蔵庫(5社)

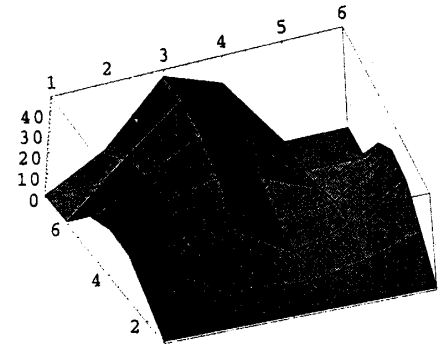

図2 ワイドTV(6社)

4 おわりに

本論文では、WWWの利用者の立場から、閲覧を継続するかあるいは中止するかを判断する材料として、サイトから得られる情報量を数量した指標を考察し、その有効性を示した。今後は、Siteの作成者の立場から凸形状をリンクの深さ(中央値、最瀕値他の量)の面から吟味し、この指標が設計指標としても適用できる可能性を検討する。

参考文献

[1] Conklin,j. Hypertext:An Introduction and survey IEEE Computer vol.20 no.9 P17-P41 1987

[2] R.E.Horn 松原訳:ハイパーテキスト情報整理学 日経BP

[3] 小池英樹:インタラクティブ3次元情報視覚化、コンピュータソフトウェア、Vol.11 No.6 P20－P31 Nov. 1994

[4]「納豆ビュー」:http://www.myo.inst.keio.ac.jp/groups/IPS/NattoView/

[5] 長尾 真編:自然言語処理、岩波書店、1998