ISS-P-132

リンク解析を用いた重要段落の抽出

冨永　拓弥　　三浦　孝夫
法政大学理工学部創生科学科

1　前書き

グラフ構造を持つデータが与えられたとき，構造の特性(節点とリンク)を利用して有用な情報を抽出する手法をリンク解析という[1]。この手法は Web 検索の有用性の他に、計量書誌学、社会学、文化人類学などで、引用解析や社会ネットワーク解析などの名前でも研究されている。この解析を通じて、重要度(節点のグラフ中の重要性)と関連度(2 節点間の関係) という観点から評価できる。本稿では、文書の段落を単位とする意味に対し、リンク解析から重要な段落を抽出する。

2　重要段落の抽出

リンク解析の最も重要な応用が重要部の検出である。Web ページでは、多数リンクされているページは重要とする。直接リンクされる部分に着目すれば定義が荒すぎ、とくにスパムに脆弱となる。様々な重要度算出法が提案され、とくに PageRank と HITS アルゴリズムが代表的である[1]。

PageRank は引用されているページが重要度を等分に伝搬するとし、全ページの重要度ベクトル **X** はリンク行列 **A** に対して $\mathbf{X} = \mathbf{A}^t\mathbf{X}$ で表される固有ベクトルに対応する。一方、HITS では、文献 a,b が同時に参照する節点の数(HITS1 得点)を $\mathbf{X}=\mathbf{A}\mathbf{A}^t\mathbf{X}$ で、同時に参照される節点の数(HITS2 得点)を $\mathbf{X}=\mathbf{A}^t\mathbf{A}\mathbf{X}$ で得る固有ベクトルに対応する。

3　リンクと類似度

文書を段落に分け、各段落は出現する自立語の出現頻度を値としてベクトルに表わす。また段落同士の関連をリンクで表現する。互いの段落の類似度（共通の単語を持つ、単語の分布が似ているなど）を方向性としてリンク解析を行う。段落同士の類似度が最も高い段落が重要段落として抽出される。本稿では４種のリンクを定義する。

(1) 内積: 文書ベクトル **d1,d2** 同士の内積**<d1,d2>** をリンク値とする。
(2) 余弦値: 文書ベクトル **d1,d2** に対して **<d1,d2>/||d1||d2|**
(3) 高頻度語のみを抽出し余弦値をリンク値とする
(4) 語の分布から得る **KL(d1||d2)** を用いる。KL 値は分布が類似するほど小さくなるため、$100 - D_{KL}$をリンク値とするが非対称である。

4　実験

以下では青空文庫から「オオカミと七匹の子ヤギ」を用いて PageRank 値、2 種の HITS 得点での５位までを表１に記す。

表１　ランキングの５位までの段落

リンク解析	PageRankアルゴリズム				HITSアルゴリズム	
類似度	内積	cosθ	高頻度語	100−KL	100−KL	100−KL
作品					権威得点	Hub得点
オオカミと七匹の子ヤギ						
重要度1位	2	6	26	1	12	11
重要度2位	25	22	27	4	14	12
重要度3位	6	2	13	12	11	4
重要度4位	22	16	12	17	17	17
重要度5位	1	24	5	11	13	15

「オオカミと七匹の子ヤギ」では、留守番を頼まれた子ヤギたちがオオカミに食べられてしまうが買い物から帰ってきた母親に生き残った一匹と子ヤギたちを救う話である。内積/PageRank アルゴリズムでは、単語数の多い段落や大きな段落を抽出している。余弦値/PageRank では、物語の要点を抽出したと考えられる。また二つが重なる抽出段落は、お母さんヤギが留守番を頼む段落、オオカミがやってくる段落、お母さんヤギが子供を救う段落である。高頻出語余弦値/PageRank では、オオカミが話す段落と母さんが帰ってくる段落といった特徴的なｍのを示している。KL/PageRank では動作や物語の前置きが抽出されている。冒頭段落や足に粉をつける段落など話題の開始部に相当する。KL/HITS では、主人公や周りの登場人物の動作の始まりや終わりの段落が抽出されている。

6　結論

様々なリンク値を定義することで、特徴的な重要段落を抽出した。とくに、高頻出語余弦値,KL ダイバージェンスで主要な働きを示した。

参考文献

[1]Amy N.Langville,CarlvD.Meyer,岩野和生,黒川利明,黒川洋,”PageRank の数理,”共立出版