人間の直感に基づく研究発表のランキングシステム

*西原陽子 [1]　砂山渡 [2]　谷内田正彦 [1]
1 大阪大学大学院基礎工学研究科 〒560-8531 大阪府豊中市待兼山町 1-3
2 広島市立大学情報科学部 〒731-3194 広島市安佐南区大塚東 3-4-1
nisihara@gmail.com

Abstract: 人間の直感に基づき研究発表のタイトルを見た目でランキングするシステムを提案する．人は自分の興味に合った研究発表しか聴講せず，興味がない分野の良い発表を聞き逃してしまう恐れがある．また，良い研究をしていてもタイトルが稚拙であるために，人に聴いてもらえないこともある．そこで人間が良いものを何となく良いと判断できる直感に見られる性質を調べ，その性質を用いて研究発表をランキングするシステムを作成した．これによって興味がない分野の良い発表の推薦を行い，発表タイトルのつけ方の方針作成を目的とする．

１．はじめに

研究発表を聴くことで研究のアイデアやサーベイを行う．
人は自分の興味にあった研究発表しか聴かない．

従来の論文推薦システムはユーザの興味をモデル化し，ユーザが興味を持ちそうな論文だけを推薦していた．中島らはユーザが論文検索を行うときに与えるクエリーと文献との関連度を，MS 被服と特定度から求め，文献をランキングする方法を提案している [中島 99]．また，山本らは既に持っている文献をクラスタリングすることでユーザの興味をモデル化し，クラスタに当てはまる新しい論文を推薦する方法を提案した [山本 05]．両手法はユーザの興味に基づく推薦であり，ニュースやブログ記事の推薦システムでもユーザの興味に合う記事の推薦を目標としている [上原子 05]．興味の対象とならないものは良いものであっても推薦されることはなかった．

自分の興味に合うものだけを選んでいては知見は深まらない．知っていることのサーベイは意味がない．分野が異なっていても良い研究発表は進んで聴講すべきである．だが，興味のない分野であるから，どれが良い発表かを判断することが難しい．

一方で人間は良いものを何となく良いと直感的に判断することができる．大辞林において直感は「推理・考察などによらず，感覚的に物事を瞬時に感じとること」とある [大辞林]．直感は敬けんに基づくものであるため，自分が経験したことがない分野の良し悪しを判断することはできない．一人の人間の直感の適用できる先は限られているが，多くの人間の直感を統合することができれば，あらゆる分野の良い悪いを判断できると考えられる．

そこで本論文では発表タイトルの印象に基づき，世の中に広まる可能性が高い，良い研究の発表を，人間の直感に基づきランキングするシステムを提案する．

２．直感の性質を調査する予備実験

評価関数の作成のために人間の直感に見られる性質を調べる予備実験を行った．本研究における直感は「多くのモノの中から良いものを選び出す基準」と定義づける．

予備実験では研究発表のタイトルを見せ，「何となく良い，面白そう」と思うタイトルを選んでもらう実験を行った．用いたタイトル集合は第２０回人工知能学会全国大会で発表される，合計２８４個のタイトルである．被験者に与えた指示は以下の通りである．

「ここから下には、研究テーマのタイトルが並んでいます。この中で「良さそう」「面白そう」と思うもの全てにチェックを入れてください．選択する時間の目安は１０分で、１５分経過したらと終了してください．全部で２８７個のタイトルがありますので、１個あたり２秒程度で判断してください。」
被験者にはタイトルをチェックしてもらった後，タイトルを選んだ自分なりの基準を記述してもらった．被験者は情報学科の大学生，大学院生５８人であった．

実験結果を述べる．被験者が選んだタイトルの平均数は 34.4 個であり，最も多く選んだ人は 190 個，最も少なく選んだ人は２個であった．選んだ人が多かったタイトルの例を表 1 に示す．表 1 においてポイントは一人の人間が選んだタイトル数 N に対し，各タイトルに 1/N を加えたときの合計を表す．表２には全く選ばれなかったタイトルを示す．

選ばれた回数が多かったタイトルはタイトル長が短いものが多く含まれていた．タイトル長は平均 24 文字であったが，その半分の 12 文字以下のタイトルは選ばれた回数の上位 20％に含まれた．また，表 1 からは「サッカー，迷路，ロボット，ボクシング，音楽」など，一般人にもなじみがあるキーワードが含まれていることが多かった．さらに表２の全く選ばれなかったタイトルを見ると．意味の分かりにくいキーワードが多く含まれていた．この結果から，分かりやすい単語が多く含まれていることとタイトル長が短いことが積極的に選ばれる基準となることが分かる．

また，タイトルを選んだ基準を書いてもらった記述文を分類したところ，表３に示す６種類に分類できた．表３では「自分の興味，関心に合う」と書いた人が最も多かった．また，「何となく面白そう」や「意味が分かる」ことも基準とした人が多かった．表 1，表２でも興味に合うキーワードが含まれるタイトルが選ばれ，難しい単語が含まれるタイトルが選ばれなかった．

以上より，多くのモノの中から良さそうなモノを選び出す基準には，自分の興味関心に合うこと，何となく面白そうと感じられること，難しい単語がなく，タイトル長が長くないことの３つがあることが分かった．個人の興味関心を表すことは難しいため，まずは何となく面白そうと感じられること，難しい単語がないことの２つの評価基準を用いて，タイトルを評価することにした．

表 1．被験者実験で選ばれた回数が多かったタイトル

タイトル	ポイント
サッカーシミュレータ環境における模倣による行動の学習	1.86
迷路問題における難易度指標の導入と実時間探索アルゴリズムの性能解析	1.47
Web 画像を手がかりとした，人物に関する情報抽出の検討	1.41
ディスプレイロボットを利用した物体の擬人化	1.29
ボクシングにおけるスキル習熟過程について	1.25
人間乱数の分析	1.24
音楽のデザイン転写技術の開発にむけて	1.16
画像情報を用いた自律移動ロボットの位置姿勢計測	1.14
心理状態認識を用いたペットロボットの行動選択手法の提案	1.11
赤外線センサーネットワークによる人物追跡	0.98

表２．全く選ばれなかったタイトル

多値議論の論理に基づく議論するエージェントシステム
文の構造と結束性に寄与する特徴的な語を考慮した文間依存関係に基づく文書要約手法の提案
ElGamal 暗号を用いた Secure_DisCSP アルゴリズムの実装と評価
OWL と SWRL を用いたロール概念の取り扱いに関する一考察
Subset-Releif 法によるデータマイニングのための属性選択手法
カニングアントを用いた ACO の構成について
形式的概念分析を用いた概念階層間の関係の発見
索引層を用いた SOM の学習高速化：初期マップ生成アルゴリズムの改良
情報編纂（Information_Compilation）の基盤技術

表３．選んだ基準に関する自由記述文の分類

選んだ基準	人数
何となく面白そう	16
自分の興味，関心に合う	40
意味が分かる	12
自分の身近に存在する	3
見た目が良い	5
役に立つ，実用的	9

３．研究発表のランキングのアイデア

本章では研究発表のランキングのアイデアを述べる．前章で直感に見られる性質を述べ，提案システムでは２つの性質を用いるとした．１つ目は何となく面白そうと感じることであるが，これはタイトル中のキーワードやその組合せの面白さを評価していると仮定する．すなわち，タイトル中のキーワードの組合せが今までにない斬新な組合せであるほど良いタイトルがつけられているとする．２つ目は難しい単語がなく，タイトル長が長くないことであるが，これは含まれるキーワードが簡単でキーワード数も少ないと仮定する．すなわち，より少ないキーワード数で，より単純なキーワードが使われているタイトルほど良しとする．

４．研究発表のランキングシステムの構成

本章では提案システムの構成を説明する．　本システムは，タイトルに含まれるキーワードの頻度と任意のキーワードの相互情報量から，タイトルに評価値を与える．システムへの入力はタイトルの集合であり，出力は評価値の降順に並べられたタイトルである．

システムでは，入力された各タイトルに含まれるキーワード（名詞のみ）を取り出し，キーワードの頻度を検索エンジン yahoo での検索ヒット数として求める．続いて各タイトルから頻度の上位３つのキーワードを取り出し，任意の２つのキーワードの共起頻度を検索の同時ヒット数として求める．続いて２つのキーワード *A, B* の相互情報量 *I(A;B)* を式 (1) で求める．

$$I(A;B)=P(A\wedge B)/P(A)P(B) \quad (1)$$

ただし，P() はキーワードの存在確率を表す．

求めたキーワード頻度と相互情報量からタイトルに評価値を与える．各タイトルには素点を与えておき，キーワード頻度がある閾値以下のキーワードの数と相互情報量の値がある閾値以下の数を素点から引き，得られた点を評価値として与える．タイトル *title* の評価値 *score* は式 (2) を用いて与える．

$$score(title)=S\text{-}key(title)\text{-}combi(title)*a \quad (2)$$

ただし，S は素点，key() は頻度が閾値以下のキーワードの数，combi は相互情報量が閾値以下の組合せの数，a は定数を表す．

式 (2) を用いて入力されたタイトルに評価値を与え，最後に評価値の高い順にランキングしたタイトル集合を出力する．

５．おわりに

本論文では直感に基づく研究発表のランキングシステムを提案した．現在は提案した式 (2) によるランキング結果と予備実験での結果を比較し，考察を行っているところである．

参考文献

[中島 99] 中島誠，金子雄一知，伊藤哲郎：用語間の概念的関係を考慮した測度による文献のランキング，信学論 (DI)，Vol.J82，No.3，pp.467 - 477，(1999).

[山本 05] 山本雅夫，矢島敬士，絹川博之：論文フィルタリング向け研究者興味の構造化表現方式，心学論 (DII)，Vol.J88，No.11，pp.2232-2245，(2005).

[上原子 05] 上原子正利，池田貴紀，浅井一希，古谷楽人，内藤裕紀，小柳滋：ニュース・ウェブログ記事集約サイトの開発，信学論 (D1)，Vol.J88，No.2，pp.305-315，(2005).

[大辞林](URL) http://www.sanseido.net/