情報の重み付き属性値の比較に基づく情報の妥当性判断支援手法

宮森 恒†
†京都産業大学コンピュータ理工学部

1　はじめに

Webコンテンツは，その量や多様性といった面で非常に膨大となっているが，その質という意味では，本当に役立つ質の高い情報と，根拠のない嘘やデマといった質の低い情報が玉石混淆の状態で存在しているという問題がある．現在，数々の検索エンジンが利用可能であり，キーワードを含む情報収集は容易となったが，一般利用者がそれら検索結果から情報の質を効果的に見分ける手段は存在しているとはいえない．

本稿では，人間の情報に対する価値基準は個々人で異なることを考慮し，信頼性が不明確な情報と，各利用者が登録した信頼できる規範となる情報とを比較することで，個々人に合わせてその情報の妥当性を効率良く判断する手法を提案する．比較においては，その情報を構成する属性と値に対し，重みをかけながら基準との差異が強調されるように非類似度を定義する．ここでは，料理レシピを例として，指定したレシピが，同じ料理の他のレシピと比べて自分にとって平均的な味付けになるのか特異な味付けになるのかを効率よく把握できることを示す．

2　提案手法

本稿では，情報 T を式1のように定義する．

$$T = \{B, C\} \tag{1}$$

where

$$B = \{l_i | i = 1, \ldots, N_b\} \tag{2}$$

$$C = \{f_i, q_i, w_i | i = 1, \ldots, N_c\} \tag{3}$$

ここで，B は分量や重みを持たない T を説明するラベルの集合であり，名称や著者，日時といった T の書誌的情報を表す．C は値 q_i と重み w_i が付随した，T を構成する属性 f_i の集合で，本稿ではこれを T の組成と呼ぶ(図1)．N_b, N_c はそれぞれ B, C の要素数である．

次に，ある比較対象 T_t と比較基準の情報 T_s を比べる際，提案法では各情報の組成 C_t, C_s を用いて比較する(図2)．T_t, T_s の非類似度 $D_{t,s}$ を式4で定義する．

$$D_{t,s} = \alpha D_c + \beta D_t + \gamma D_s \tag{4}$$

ここで，$\alpha, \beta, \gamma \geq 0$ であり，図2および式4で，D_c は，T_t と T_s に共通して含まれる属性による非類似度を表し，D_t, D_s は，それぞれ T_t, T_s にのみ含まれる属性を用いた非類似度に対応する．特に，D_t, D_s は，それぞれ比較対象，比較基準にしか含まれない属性を用いているが，例えば，D_t は，比較対象の全属性数に対し，比較対象にのみ含まれる属性数がどの程度の割合で含まれるかに基づく関数で定義したり，D_s は，比較基準にのみ含まれる属性の重みに基づく関数で定義される．これにより，比較対象の固有性や，比較基準に含まれる重要な属性をどの程度含んでいないかといった側面を強調することができる．

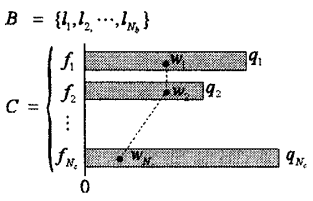

図1: 情報 T の定義

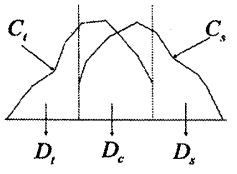

図2: 比較対象 T_t と比較基準 T_s との比較

従来，組成 C のような形式をもつ集合は1つのベクトルと考えることもでき，コサイン類似度で2つのベクトルの類似性を測ることができる．しかし，この方法では，例えば，レシピの場合，さまざまな種類のたくさんの料理から，特定のパスタ料理を識別するには有効だが，同じ種類のパスタ料理の中で，これは味が濃い，うすいといった識別をするのは困難であった．

本提案手法により，特定のパスタ料理での味の濃い/薄いといった違いを，基準との差を強調することで比較でき，より的確な判別につながることが期待される．

3　料理レシピへの応用

料理レシピを例として提案手法を適用する．料理レシピは，一般に，タイトルや作者等の書誌的情報，材料一覧，作り方手順から構成されている．本稿では，書誌的情報および材料一覧を用いた．材料の名前や分量を，情報の属性名，値と考え，提案手法を適用した．

実験データは，料理の講師と一般投稿者によるレシピが利用可能な“みんなのきょうの料理”[1]から約7千件，一般投稿者によるレシピが利用可能な“クックパッド”[2]から約43万件のレシピを用いた．

書誌的情報は，あらかじめ確認したdivタグ属性等を参照することで検出する．また，材料一覧については，いくつかの統語パターンで分量と単位を検出し，その

Assisting the Validity Assessment of Information based on Composition Similarity

† Hisashi MIYAMORI (miya@cse.kyoto-su.ac.jp)
Faculty of Computer Science and Engineering, Kyoto Sangyo University (†)

検出位置から材料名と備考を抽出する．さらに，表記ゆれに対処するため，省略形の補完や，単位変換，分量の正規化等を施す．

提案手法によるレシピの妥当性判断を支援するプロトタイプシステム「味コレ！」を実装した (画面は割愛)．調べたいレシピと，利用者が選択した基準 (例えば，濃い味好きの人の基準等) とを比較することができ，その結果として，縦軸に各材料，横軸に，同じ料理の他のレシピの平均を1としたときの，対象レシピの各材料の分量の割合を示すグラフを提示する．これにより，一見しただけでは判断が難しいレシピの味付けの具合を，自分の基準に合わせて判断しやすくなる．

4 実験と考察

まず，インデキシングによる材料名，分量，単位の各検出精度について調べた．“きょうの料理” および “クックパッド” からランダムに各100件のレシピを選び，材料名，分量，単位について検出精度を集計した．材料，分量，単位ともそれぞれF値で0.944, 0.961, 0.950となり，概ね高い精度で検出できていることを確認した．

次に，コサイン類似度を使った従来手法と提案手法との比較を行った (図3,4)．図の各ID文字列は1件のレシピに対応しており，ここでは約400件の酢豚を多次元尺度法によって2次元上に配置したものである．横軸は濃い味が好きな人との非類似度，縦軸はうす味好きな人との非類似度に基づいた結果を表している．コサイン類似度では，酢豚レシピの分布がある部分に集中して得られており，同じ酢豚という料理の中での特徴を判別しにくいことが分かる．一方，提案手法では，基準との差異が強調されたことにより，濃い味好きの人が好きなレシピは薄味好きの人はあまり好きではないという反比例の関係を示す分布が得られており，より利用者基準に合致した特徴を分離しやすくなっていることが確認できる．

5 関連研究

情報の信頼性については，Webページのどのような要素が利用者のそのページに対する信頼度に影響を与えるかについて大規模な被験者実験と分析結果が報告されている [3]．また，利用者に，発信者や社会的評価等の多面的な分析結果を提示することで，その情報の信頼性判断を支援するといった研究がなされている [4][5]．

本手法は，利用者の価値基準の違いを考慮し，各自が自分の基準に照らしてその情報の妥当性を的確に判断しやすくできる点で従来手法と異なっている．

6 まとめ

利用者が登録した規範となる情報と，信頼性が不明確な情報とを比較することで，情報の妥当性を効率良く判断する手法を提案した．料理レシピへ応用したシステム「味コレ！」を試作し，与えられたレシピが，個々の利用者の基準に合わせて効率よく判断できるようになることを確認した．

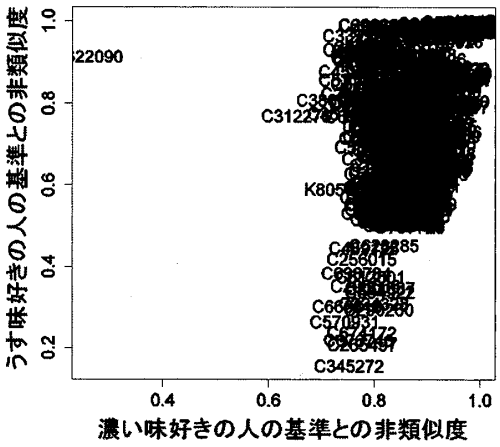

図 3: コサイン類似度に基づく比較

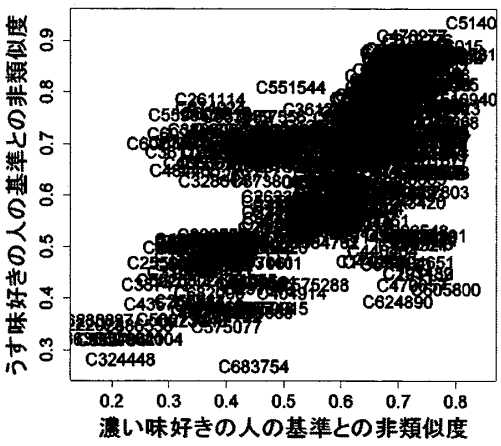

図 4: 提案手法に基づく比較

謝辞

本研究は科研費基盤 (B)20300042，および，京都産業大学総合研究支援制度の助成を受けたものである．

参考文献

[1] みんなのきょうの料理, http://www.kyounoryouri.jp/

[2] クックパッド, http://cookpad.com/

[3] B. J. Fogg, et. al.: What makes Web sites credible?, ACM SIGCHI conference on Human factors in computing systems, pp. 61–68, 2001.

[4] 田中克己: サーチエンジンにおける信用度・品質評価について, データベースとWeb情報システムに関するシンポジウム (DBWeb), pp.107–108, 2006.

[5] 黒橋禎夫: 構造的言語処理による情報分析研究, 言語処理学会第13回年次大会 (NLP2007) ワークショップ (W2), pp.17–18, 2007.