4N－5

知識管理のためのテキストマイニング

武田 浩一 (takeda@trl.ibm.co.jp)
日本アイ・ビー・エム株式会社 東京基礎研究所

1　まえがき

大量のテキストデータから新たな知見や規則性を獲得するテキストマイニング [8] は、組織における知識資産を効果的に利用するための知識管理 [6] の手法においても重要な要素になるものと考えられる。ここで研究として興味深いのは、今までのテキスト処理 (例えば検索) 手法とテキストマイニング手法とが以下の点でテキストに対して決定的に異なるアプローチをとることである。

- テキスト検索、クラスタリング、分類といったテキスト処理手法では名詞句のような客観性の高いデータを元に文書へのインデキシング、類似度判定などを行い検索結果を返す。
- テキストマイニングでは、名詞と動詞との係り受けといったより複雑な構造をもつ情報や、主観的な情報表現 (助動詞や副詞) が文書の分類や獲得すべき知見として重視される [7]。

例えば、

「PCをインターネットに接続することができない。内蔵モデムは正しく動作しているのか？マニュアルの記述はわかりやすくしてほしい。」

というテキストと、

「PCをインターネットに接続することができた。内蔵モデムは正しく動作していた。マニュアルの記述はわかりやすかった。」

という２つのテキストは、従来のテキスト検索の観点からはほぼ同一のインデクスをもち、検索式に対する類似度もほとんど差がないと考えられる。テキストマイニングにおいては、これらのテキストは非常に異なる情報をもつ。前者では否定的情報 (不満)、質問、要望といった情報が含まれるのに対して、後者ではより中立的な記述および満足を示す情報が含まれているものと判定される。さらに前者のようなテキストが大量に存在する場合は、記述されているPCか内蔵モデムに問題がありそうで、マニュアルの記述は改善を要することが結論づけられる。また、より深いマイニングによって、インターネット接続やモデムに関する不満や疑問と、マニュアルの記述の不備との相関関係を明らかにすることができる。

2　１次資料と注釈

知識管理においては、知識とは単なる情報とそれを活用するコンテキストや意思決定を記述したメタ情報を含むものと考えられる。我々がこのような情報を形式的に記述するためには、述語論理やRDF(Resource Definition Framework)[4] といった記述様式を必要とするが、日常的にはこれらすべての情報がフラットなテキストによって表現されている。ただし深い言語処理によってこれらの識別を行うことは技術的にまだ困難であると考えられるため、ここではより現実的なものとして知識資産を構成する１次資料と、協調的なオーサリングなどでよく利用される「注釈 (annotation)」をメタ情報として用いる以下のような状況を仮定する。

- 知識管理において利用される１次資料ｅの集合をＲとする。
- １次資料ｅまたはその部分的要素ｆに対して、テキストで書かれた注釈ｔを対応づける。

具体的には、１次資料e(あるいはf) および注釈ｔはともにXML[1] で表現し、その対応は例えばXPointer[2] で記述するといった実現方法を想定している [3]。前者はまた、HTMLで書かれたWebページであってもよい。

これにより、１次資料に対するユーザの補足や評価、関連したコンテキスト、それに基づいて下された決定事項といった情報が柔軟に付与できる。従来から知識管理におけては知識の入力が大きなボトルネックの１つであることが指摘されているが、この方式では

- １次資料に対して少ない労力で実用的で有用な情報が付与できる。
- 付与された情報をもとに、より効果的な知識共有やさらなる情報の付与が促進される。
- 注釈の付与により情報フィルタリング [5] 的に１次資料の有用性を測定する尺度が得られる。

Text Mining for Knowledge Management
Koichi Takeda
Tokyo Research Laboratory, IBM Research

といったメリットを実現できる。注釈としては、「データのサイズが大きいときにはうまくいかない」、「本手法の精度が高く提案システムに採用した」といった自由なテキスト記述を用いる。

3 テキストマイニング

1次資料と注釈とは、それぞれを節点と考えることで2分グラフを形成する。注釈には事前に定義された種類(注意、満足、質問など)をもってもよい、この場合にはラベルつきの枝により表現できる。この2分グラフとテキスト記述を元に以下のような処理が可能になる。

- 1次資料のランキング：可能や満足を示す注釈の数によって順序付けが可能である。問題や注意を示す注釈の数は、否定的な重みとして利用できる。
- 1次資料のクラスタリング：通常のテキストマイニングによる1次資料のクラスタリングだけでなく、注釈に現れる単語や句によってそれに対応する1次資料のクラスタリングを行う。1次資料のカテゴリ分類についても同様である。
- 1次資料と注釈との相関規則：1次資料の部分要素に対して注釈がつけられているときには、部分要素をおよび注釈を統語的あるいは概念的に同値分類し、両者の相関規則を獲得できる。

特定の1次資料については、その部分要素に対する注釈全体が対応すると考えることで集約化が行える。その際に、評価が有用か問題ありといういずれかに集中していればその資料の価値はあまり矛盾なく決定でき、両者が等分になるようであればその内容について、検討あるいは修正が必要であることがわかる。また、注釈に対する注釈を許せば(モデリングも複雑にはなるが)より高次の知識を扱うことが可能となる。

4 知識創造のための機構

知識管理および知識再生産のループの中にテキストマイニングがはいることにより、注釈を利用して知識に対する需要と供給の把握を行うことができる。この事実は、コールセンターにおける問い合わせ数やその内容の変化からも経験的に知られているが、例えば、新たな製品が発表されると急速にその製品に対する問い合わせが増加し、それが利用可能になった場合にはその利用法についての問い合わせ、トラブルの報告、評価といった情報が集まるようになる。タイムリーにFAQを提供し、正しい情報を提供するとともに、その製品についてのそれまでのすべての情報要求と関連情報とを対応づけることでひとまとまりの知識が形成される。上記の1次資料と注釈のモデルでいえば、

- ある製品発表に相当する1次資料が生成されると、それについて詳細を要求する注釈が単調に増加する。
- 詳細情報があきらかになれば、その1次資料に対して「詳細はこちら」という注釈が付与され、詳細要求の注釈数の増加が停止する。
- 製品の普及により新たな1次資料や注釈が発生し、その注釈の分布によって改良、広告、企画といったアクションをともなう意思決定が行われる。新たな情報が1次資料に付与される。

といったフェーズをへる。直観的には、問い合わせ過多や増加の傾向は知識の需要不足であり、問題の指摘が増加する場合は相当する対応をせまられる状況を示す。テキストマイニングは、このような問い合わせや問題の指摘の内容や傾向を統計的に把握し、適切なアクションをとるために必須の技術となる。

5 今後の課題

本論文で提案した1次資料と注釈による知識管理は、インターネットのnewsgroupで見られる質問やコメントとその返答のような方式で運営できるものと考えている。膨大な量の情報に対応するには、さらに情報の可視化技術や商用のグラフ表現を利用した知識管理システムの技術を統合する必要がある。2分グラフを利用したテキストマイニングのより多様なアルゴリズムも今後の課題である。

参考文献

[1] The World Wide Web Consortium. "Extensible Markup Language (XML) 1.0. W3C Recommendation". http://www.w3.org/TR/REC-xml, Feb. 1998.

[2] The World Wide Web Consortium. "XML Pointer Language (XPointer). W3C Working Draft". http://www.w3.org/TR/WD-xptr, Mar. 1998.

[3] The World Wide Web Consortium. "Annotation of Web Content for Transcoding. W3C Note". http://www.w3.org/TR/annot/, July 1999.

[4] The World Wide Web Consortium. "Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation". http://www.w3.org/TR/REC-rdf-syntax/, Feb, 1999.

[5] U. Shardanand and P. Maes. "Social Information Filtering: Algorithms for Automating "Word of Mouse"". In *Proc. of CHI'95*, 1995.

[6] 紺野. "知識資産の経営", ISBN4-532-14636-4. 日本経済新聞社, 1998年.

[7] 諸橋, 那須川, 長野. "テキストマイニング：膨大な文書データからの知識獲得－意図の認識－". 情報処理学会第57回全国大会 5K-3, 1998年.

[8] 那須川, 諸橋, 長野. "テキストマイニング－膨大な文書データの自動分析による知識発見". 情報処理, 40(4):358–364, 1999年4月.