自然言語処理 129－9
（1999. 1. 21）

インタラクションを重視したテキスト情報操作環境

嶋田 敦夫　藤田 克彦
株式会社リコー　研究開発本部
shimada@rdc.ricoh.co.jp, katsuhiko.fujita@nts.ricoh.co.jp

本稿では、ドキュメント空間に対するインタラクションのモデル、Document Spreadsheet model を提案する。

我々の提案するDocumentSpreadsheetモデルの基本的な考え方は、ドキュメント集合に対して導入された種々のメトリックを有効活用するために、それから得られた数値、個々のドキュメント、メタデータなどの様々なデータを統一的に扱うインタフェース層を用意し、利用者による処理の制御や情報間の関係把握を容易にすることである。

このモデルを実装し、テキストで記述された顧客窓口への問い合わせデータに対する分析作業に応用し、モデルの効果を確認した。

An Interaction Model for Document Spaces

Atsuo Shimada and Katsuhiko Fujita
Research and Development Center, RICOH Co., Ltd.
shimada@rdc.ricoh.co.jp, katsuhiko.fujita@nts.ricoh.co.jp

A new interaction model for document spaces --- Document Spreadsheet model --- is proposed, where through tabular interface one can view and manipulate various values including metric related data and meta-data assigned to each document in an integrated spreadsheet-like manner.

Today, a lot of systems that treat document collections introduce some kind of metric for their document spaces: in document retrieval, those metric are usually used for calculating similarities between search queries and documents. But seldom those values are open to be used by users, so the effect of introducing metric is very limited.

To fully utilize those metric, interface for Document Spreadsheet model provides users with methods for manipulating data and specifying metric used and operations such as cross-tabulation, etc. Thus users can do various analyses iteratively through this interface until he get his own desired result.

1. はじめに

テキスト情報に対してメトリックを導入し、統計的な手法を適用することで、膨大なドキュメント集合を扱う試みがなされている。類似検索や検索結果のランキング、あるいはテキスト・データ・マイニングなどの技術の背景には、ドキュメントの長さや含まれる単語の出現頻度、ドキュメント

間の類似度といった様々なメトリックと、それに基づく演算がある。

例えば、検索システムでは、検索語に対して関連度の高いドキュメントを上位に表示させるため、単語の頻度やそれを何らかのモデルに基づいて変換した数値が、検索語に対する各ドキュメントの重要度や関連度を決定する情報として利用されている。情報の可視化やドキュメント・クラスタリングなどでは、ドキュメント集合をもとに空間およびメトリックを導入している。空間内に個々のドキュメントが布置されることで距離計算などが行え、類似するドキュメントのグループ化が可能になるという考え方である。

テキストに関するメトリックは、その有効性により他にも様々な場面で利用されている。個々の数値がドキュメント集合全体の特徴を表現すると同時に、またそれぞれのドキュメントに固有の、全体からの偏りも表しているからである。

我々は、ドキュメント集合の様々な側面に着目し、その集合の特性や情報間の関連性を発見するというインタラクティブな情報構造把握に、メトリック導入の効果を有効に機能させたいと考えている。そのために、本稿ではまず現状のメトリック利用に関する問題点を考察する。次にドキュメントに関するメトリックを有効に扱うための枠組みを提案し、その枠組みにしたがって開発した実装例について報告する。

2. 現状のシステムにおけるメトリック利用の問題点

2.1. メトリックと対象の選択の限定

メトリックには、ドキュメント中の特定単語の有無や、出現頻度、テキストの長さなどの他のデータに依存しない絶対的なものと、規準化測度やベクトル空間モデルなどのドキュメントが属するデータ集合に依存して相対的に決定されるものとの区別がある。前者は、他のデータに依存せず独立に決定することができる。あるドキュメントに関する任意の単語の有無やその頻度などは、そのドキュメントを異なる集合に移動させても値は変化しない。属する集合に依存して決定されるメトリックは、属する集合が変わると同じ値を取ることが保証できなくなる。例えば、ドキュメントAとドキュメントBとの類似度はそれらが属する集合において計算されるので、2つのドキュメントを別の集合に移動させると異なる類似度が算出される。

ところで、このようにメトリックには様々な種類があるので、同じドキュメント集合に対しても様々な観点からメトリックを導入することが可能である。目的に応じてそれらを切り替えて利用できるシステムは現状少ない。このため用途に応じてシステムを使い分けなければならなくなっている。

同じメトリックを採用する場合でも、その対象をドキュメント集合全体にするか、その一部にするか指定したい場合があるが、多くのシステムでは予め設定された対象についての処理だけが許されている。

2.2. 結果の比較手段の限定

メトリックによる処理結果を複数回にわたって保持し、表示する手段が提供されていないシステムが多い。例えば検索の度に画面が上書きされ前の情報が消える。そのため、ある検索語に対する結果と別の検索語に対する結果とを比較し、検索語の有効度を判断するなどが困難になる。

ドキュメント集合の可視化においても同様のことが指摘できる。従来の可視化技術の強調点は、認知的負荷の高い情報把握作業を、より負荷の低い知覚課題へ変換すること[Rao, R. et. al.,1995]であった。そのため内部のメトリックは知覚的判断を容易にするよう情報をまとめ視覚化することに使われ、複数回の結果を比較する手段が提供されているケースは少ない。可視化の場合、表示できる次元数が限定されているため、ますます比較が困難になっている。

2.3. 情報リダクションによる問題点

Schutze, H. and Silverstein,C. [1997]の報告のように、計算コストの面から、類似度計算に用いる次

元は、ある規準によりリダクションされることが多い。例えば、LSI（Latent Semantic Indexing）という特徴次元抽出方法を次元リダクションに用いれば、得られる特異値の大きさの上位100次元程度で各ドキュメントの類似度を計算することになる。このようなクラスタリングはドキュメント集合全体を大きな粒度で把握するには適しているが、この次元の限定がシステムにより予め定められ、それ以下の次元に利用者の望むクラスタ粒度がある場合、目的の結果が得られないことになる。これはシステムの高速化に対する要求を優先し情報のリダクションを行っているというやむを得ない事情から来ていると考えられるが、速度を優先するか、クラスタ粒度の自由度を優先するかは、利用者が選択できるようにすべきであろう。

2.4. 制御のための情報の不足

メトリックは通常システムによって隠蔽されている。数値そのものは目に触れることがなく、したがって利用者が操作することもできない。ところが、目的に適したドキュメント・クラスタリング結果を得ようとする場合などにおいては、例えば出現頻度に着目してどのような単語を重視するか、またどの単語をストップ・ワードに指定するかの判断が重要になる。

また、全単語を用いずドキュメント毎に高頻度の単語を一定の数だけ残す（local truncation）ことで精度を落とさず速度の向上が可能だという報告もある[Schutze, H. and Silverstein,C.1997]。この場合もどの単語を残すかが重要になってくる。

こうした判断は、利用者の利用目的や利用者のおかれている状況に依存するため、それらを利用者自身が判断するために、役立つ情報が提供されなければならない。しかしこうした情報を提供するシステムは少ない。

2.4. 他の情報との関連に関する問題点

ドキュメントに様々な情報が付随していることがある。これらは書誌事項ないしメタデータと呼ばれ、作成者や情報源、日時などが個々のドキュメント毎に付与されている。アンケート調査などで得られるフリーテキストデータ、あるいは企業の顧客相談窓口に寄せられる質問の記録などでは、この付随情報が計画的に採取されている。一方、メトリックは、ドキュメントのテキスト部分に対して利用されることが多く、そうした数値が他のメタデータなどとの関連づけて利用できる環境は少ない。

例えば、Nowell,L.T. et. al. [1996]は、Envisionという論文検索システムを開発するにあたり、集中的な利用者調査を行った。調査によれば利用者は検索に関するニーズの他に、ある専門領域で重要なトピックスのトレンドや引用の多い重要な論文の特定、研究コミュニティーの把握などをニーズとして挙げたと言う。これらの分析は、論文の発行年、著者名、引用文献リストなどの情報を、個々のドキュメントに対するメトリックと関連付けることで初めて可能となる。この例では、ドキュメントに対して算出された値と著者名や発行年のクロス集計をとることが必要になる。

3. メトリック利用のための条件

前節までの問題点を解消するには、ドキュメント集合に対するメトリックやメタデータに関して、少なくとも以下の条件を満たす統一的な操作環境が必要となる。

- 目的に応じて使用するメトリックが選択できること（例:ベクトル空間モデルやLSI空間モデルの選択が必要に応じてできる）
  - その際、メトリックを導入する対象の範囲を指定できること（例:ドキュメント集合全体、条件を満足するその一部の範囲など、あるいはメトリック構成で用いる単語を指定・選択することができる）
  - なお、算出された値の表示と、それに対する後の操作が可能なこと（例:各ドキュメントに対応させて空間での座標値を表示し、その座標値を利用者が操作できる）

- 表示された値に対して目的に応じた処理が指

定できること（例:表示されている座標値の第1次元の値でソートすることができる）

- その際、処理対象の範囲を指定できること（例:ソートの範囲が指定できる）
- なお、処理結果の表示と、それに対する後の操作が可能なこと（例:ソートした結果を表示し、上位100個の平均をとる）

・他のメトリックにより算出された値と対応させて比較・相関をとることが可能なこと（例:相関係数やクロス集計を取ることができる）

・他の情報（メタデータなど）との比較・相関をとることが可能なこと

これらのことが同一のインタフェースから繰り返し実行でき、それらを統一的に見られることが必要である。個々の条件を見れば、従来のシステムでも実現されているが、統合的に扱えない限り前述の問題を解決できないからである。

次節で、こうした条件を満たす操作環境のモデルを提案する。

4. Document Spreadsheet モデル

我々の提案するDocumentSpreadsheetモデルは、ドキュメントに関するメトリックをより自由に扱うため、表形式のインタフェース層、メトリック、ドキュメント集合層から構成される枠組みである（図1）。インタフェース層では、メトリックにより算出される数値やドキュメント、ドキュメントに付随する情報などを統一的に扱うことができる。またメトリックの選択やドキュメント集合の操作が行える。

Document Spreadsheetモデルの基本的な考え方は、データマニピュレーションのためのインターフェース層を明確に分離し、それを介してメトリックやデータを操作できるようにすることで、利用者による処理の制御やインタラクティブな情報間の関係把握を可能にすることである。

document collection

データとなるドキュメントの集合。集合のどの部分を対象とするか選択することができる。

metric

ドキュメント集合に対して定義されたメトリック。単語やドキュメントの間の距離、任意の次元への射影などを求めるオペレータを定義することもできる。

tabular interface

2次元にオブジェクトを表示し、それに対する操作を行うためのユーザーインタフェースで、オブジェクトを表示した2次元配列のセルにより構成されている（インタフェース層で操作可能な対象をオブジェクトと呼ぶ）。

このインタフェース上には、ドキュメントや単語、メタデータと、それらに対応しオペレータにより演算された結果が、オブジェクトとして配置される。

利用者からの操作（manipulation）には次のようなものがある。

ドキュメント集合層に対する操作があるが、これはメトリックの導入対象の範囲の指定を行うためのものである。

利用するメトリックとオペレータの選択・指示も用意する。我々の枠組みでは様々なメトリッ

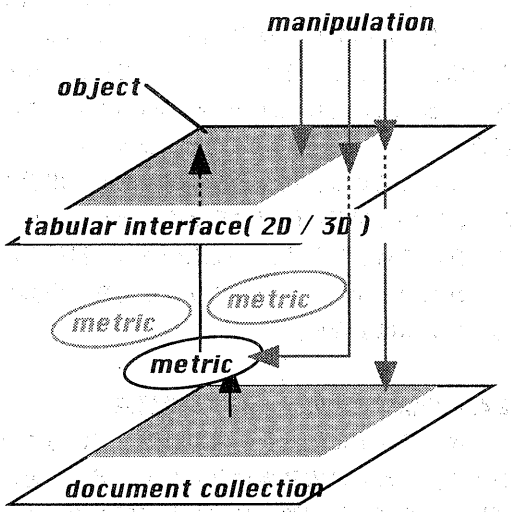

図1. Document Spreadsheet モデル

クを扱うため、利用するメトリックの選択と、そのメトリック導入によって得られた値に対する有効なオペレータの選択が必要となる。

その他、情報を扱う一般的な操作として、オブジェクトを表示したセルや行、列、あるいは行列（これらもオブジェクト）に対する各種操作が用意されている。これらは比較や相関をみるために行われる。例えばソートや、相関の演算などであるが、これらは従来からあるスプレッドシートの標準的な機能である。

Document Spreadsheet モデルによる効果として期待できるものには次のようなものがある。

まずメトリックの導入とそれにより算出された結果を利用者に直接呈示することで、ドキュメント集合における個々のドキュメントの位置づけが数値で把握できる。また複数のメトリックによる数値を比較することでそれぞれのメトリックの有効性が判断できる。

ドキュメントに付与されたメタデータと数値との相関や比較分析の機能をインターフェース上に用意したことで、従来のシステムでは困難だった両者の多様な関係を把握することが可能になる。

そしてこのようなことがインタラクティブに、かつ繰り返して同一の環境から可能になることにより、結果として、利用者は目的とする分析結果を制御することができるようになる。

5. Document Spreadsheet モデルの実装例

Document Spreadsheet モデルを実装した例として、我々は図2に示す構成によるテキスト分類システムを開発した。

各ドキュメントに対応する情報を行に配置し、各列にメトリック導入により得られた値や各種のメタデータを格納するよう構成する。行単位での選択がドキュメントの選択になり、列単位での選択が各種の値やメタデータの選択に対応することになる。

重要なメトリックとしては、ベクトル空間モデルとLSIモデルに対応する二つを用意した。LSIモデルは、単語×ドキュメント行列に対して特異値分解を施すことで、クラスタリング時に計算されるドキュメント間の類似度の算出コストを低減させる効果を持つモデルである。これらを適用することで生成された空間において、ドキュメントは座標値を持つことになる。

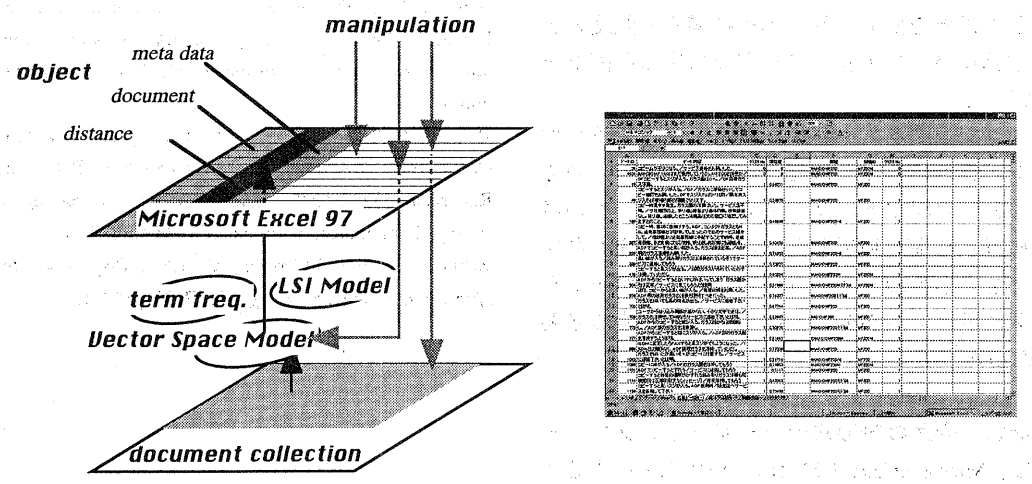

図2. Document Spreadsheet モデルの実装例（構成と画面例）

このような空間に対する座標値、これから算出できる空間内におけるドキュメント間の距離（類似度）などの他に、ドキュメント集合の分析においてしばしば用いられ有効とされるドキュメント毎の文字数や単語の出現頻度、その単語を含むドキュメント数など様々な値を算出する処理も用意し、それらを実行させる指示をインタフェースから利用できるようにした。

今回のシステムが扱う対象とするドキュメント集合の一つは、顧客窓口への問い合わせの記録である。このドキュメント集合の分析においては、問い合わせの内容の分類、それらの構成比や機種毎の差異、発売後からの問い合わせの種類の経時変化などを知る必要がある。

これらを求めるために、まず問い合わせの内容の分類が必要だが、上記の空間とそのメトリックを用いて、クラスタリングが行える。この結果をセルに表示するだけでなく、それぞれのクラスタに含まれるドキュメントの数の集計や、これらと他のメタデータ（機種名）とのクロス集計をとることが可能で、機種毎の問い合わせの内容の分布の違いなどを明らかにすることもできる。こうしたことは、同一のインタフェースから一連の作業の繰り返しの流れとして実行できる。この流れの中で、クラスタリングの対象となるドキュメントの選択を変えたり、単語に関する頻度の表示を見ながらストップワードの指定の判断なども行える。

これは、クラスタリング結果を表示するだけのシステムにはない機能である。これらと同等のことを実現するには、従来であれば、その度の演算の結果を利用者自身が管理し、様々なツールを組み合わせたり（データのインポートやエクスポートが必要）、切り替えたりする必要があった。

なお、今回の表形式のユーザーインタフェースは、Microsoft Excel 97上に、VBAなどによるアドインとして実現し、演算もこのインタフェースから実行するようになっている。

6. 今後の展開

従来分離されていたテキスト系の分析の種種のフェーズの統合を、今回のモデルでは提案した。今後は、作業者の情報を分析する過程の研究と連携して、情報分析タスクに必要となるマニピュレーションのためのプリミティブやそれらに対応する有効なオペレーション、メトリックを実装していく。

なお、このモデルの簡単な応用例としては、検索結果リストに基づきドキュメントの本文をシステムに取り込む仕組みを用意し、Document Spreadsheetモデルで扱うことで、検索結果の有効利用が図れるようにする方向もある。従来の「曖昧検索」や「類似検索」では、ドキュメントを見るだけであったので、その内容の判断は一つ一つの内容を見ながら人間が行わなければならなかった。これに比して、Document Spreadsheet モデルを応用すれば、より詳細な観点でのドキュメントの整理、活用が可能になる。

参考文献

- Hearst, M. (to appear in 1999). *User interfaces and visualization,* In Baeza-Yates,R. and Riberio-Neto,B.(Eds.), Modern information retrieval. Addison-Wesley.
- Nowell, L.T., France, R.K., Hix,D., Heath, L.S. and Fox,E. 1996. *Visualizing search results: some alternatives to query document similarity.* SIGIR-96, pp67-75.
- Rao, R., Pedersen, J.O., Hearst, M.A. , Mackinlay, J.D., Card, S.K., Masinter, L., Halvorsenm P.K. and Robertson, G.G. 1995, *Rich interaction in the digital library.* Communication of the ACM 38,4,pp29-39
- Schutze, H. and Silverstein, C. 1997, *Projections for efficient document clustering,* SIGIR-97, pp74-81.