E-008

情報検索システム"CoreExplorer"を用いたメーリングリスト間の関連トピック分析
Analysis of Associated Topics in Mailing Lists
using Information Retrieval System "CoreExplorer"

塚原 朋哉[†]　**原島 一郎**[‡]　**佐藤 俊也**[†]
TSUKAHARA Tomoya　HARASHIMA Ichiro　SATO Shunya

１．まえがき

コールセンターに寄せられる顧客の声の分析，保守日報からの不良の早期発見など，企業内部の大量の文書（テキスト）から分析者にとって有用な情報を抽出する分析技術が求められている．

しかし，有用な情報が抽出されたかどうかの判断は分析対象や分析者の観点に拠って異なる．分析対象を柔軟に絞り込みながら対象データを様々な観点で簡単に分析できるシステムを求める声は多い．

本報告では，検索結果として，文書・文書の属性・自動抽出された文書を特徴付ける特徴語を出力する情報検索システム[1]（以下"CoreExplorer"と記す）を利用したテキストマイニングを行う分析システムについて報告する．

作成したシステムを複数 ML の分析に用い，関連したトピックを持つ ML 同士の発見やトピックに関わる人の抽出が簡単にできることを示す．

２．分析システム構成

分析システムは次の３つの構成要素からなる．

- テキスト検索エンジン
- テキスト分析エンジン
- テキスト分析クライアント

以下，これらについて説明する．

2.1　テキスト検索エンジン

テキスト検索エンジンには，文書の本文中の単語や属性を利用して分析対象を細かく指定したり，様々な観点で情報を抽出することができる"CoreExplorer"を用いた．"CoreExplorer"が分析対象から抽出する情報（検索結果）は以下である．

(1) 特徴語一覧
(2) 文書一覧
(3) 文書間の関連
(4) 特徴語を含む文書一覧

ここで，出力する特徴語としては，文書中の特徴的な単語のほかに，文書の属性を指定することも可能である．また，"CoreExplorer"では属性を等価的に扱うことが可能なため，アンケートの分析では属性として年齢・性別を，保守日報では機器名・顧客などと属性を自由に設定することができる．また，抽出する属性の指定も自由に行える．これにより，分析者が求める観点（属性・文書中の単語）を自由に変更しながら情報を取得することが可能となる．

2.2　テキスト分析エンジン

テキスト分析エンジンでは分析条件に従い検索エンジンに検索要求をし，2.1 節(1)〜(4)の情報を取得する．月ごとの傾向分析などでは検索エンジンに複数回検索の要求を投げる．

検索エンジンから取得した情報を元に以下の情報を生成する．

(5) 文書のグループ（(3)文書間の関連から）
(6) 特徴語間の関連（(4)特徴語を含む文書一覧から）
(7) 特徴語のグループ（(6)特徴語間の関連から）
(8) 文書に含まれる特徴語一覧（(4)特徴語を含む文書一覧から）

テキスト検索エンジンから得られる(1)〜(4)の情報と，テキスト分析エンジンで生成する(5)〜(8)の情報により分析データを整形し，その結果を 2.3 節のテキスト分析クライアントに出力する．

(5)文書のグループ，(7)特徴語のグループの作成には，最小距離法による階層的なクラスタリングを行った．その際の距離としては，(5)では文書中の単語の重要度をベクトルの要素とする文書ベクトルを用い，ベクトル同士のコサイン角を用いた．(7)の距離は単語ベクトル（単語を含む文書中の単語の重要度を要素とするベクトル）により(5)と同様に算出した．

2.3　テキスト分析クライアント

テキスト分析クライアントは，分析者の要求を受け付け，要求をテキスト分析エンジンに送信し，分析結果を受け取り表示する．

表示する情報は 2.1 節テキスト検索エンジン, 2.2 節テキスト分析エンジンで生成した情報のうち，特徴語に関しては(1), (4), (7)，文書に関しては(2), (5), (8)の情報である．

(5)特徴語グループや(7)文書グループの階層的なクラスタを入れ子構造の表形式で表現した（図１中上・右上）．

(4)特徴語を含む文書一覧や(8)文書に含まれる特徴語一覧は(5)や(7)のグループを単位として表示することも可能であり，文書や関連文書グループにどんなトピックがあるのか，あるトピックを持つ文書グループはどれか，特徴語と文書を相互に調べることが可能である．

2.1 節で出力する特徴語を，文書中の重要語とすればトピックを，人とすればトピックに関わる人を分析することができる．

月ごとの特徴語の推移を見ることでトピックの移り変わりに気付くことも可能である．

[†] 株式会社日立東日本ソリューションズ
[‡] 株式会社日立製作所　日立研究所

図 1　分析結果（ML 分析例）

3．分析システムの適用と評価

本報告で作成した分析システムを複数 ML の分析に適用した．一般に ML の分析では，「発言者の偏り」や「月当たりのメールトラフィック」などは分析できるもの，記述された内容そのものの分析にはテキストマイニングの技術が必要となる．今回の適用先では社内で 100 近くある ML のうち，分析可能な 40 の ML に対し，分析作業を行った．

3.1　分析準備

分析の前処理として，それぞれの ML から次の３種類の分析対象データを作成した．

(i)「メール単位」：メール１件を１文書としたデータ．メール中から【本文】，【送信者】，【送信日】，【タイトル（サブジェクト）】を抽出した．

(ii)「月単位」：メール一月分を１文書としたデータ．一月分のメールの本文を連結して【本文】とし，【送信者】としてその月にメールを送信した人全員を，【送信日】を最新の送信日に，【タイトル】を ML 名＋年月にしたデータを作成した．

(iii)「ML 単位」：ML 内全てのメール本文を結合して１文書としたデータ．メール量が多く本文が長くなる場合は，最新のメールから一定量を制限として【本文】を抽出した．

3.2　分析処理

分析対象の絞込みには，メール本文中の単語や，【送信者】や【送信日】などの属性が利用できる．出力特徴語としては，メール本文中の単語（トピック）や，送信者などの属性を指定できる．

3.1(ii),(iii)の月単位・ML 単位で分析したところ，文書のグループ表示（図 1 右上に相当）から共通の話題を持つ ML 同士がグループ化されることが確認された．一例として，複数の ML にクロスポストされているメールがあると，そのメールを含む ML 同士が同一のグループや近いグループに分類された．

これらのグループを選択すると，ML のグループとグループ内に含まれる特徴語の対応が表形式で表示されるため（図 1 右下に相当），グループ内の文書がどのような内容で関連しているかを把握できる．出力する特徴語として送信者を選択すれば，関連する ML に関わっている送信者一覧を知ることができ，同様のトピックを持つ ML で発言する送信者のグループを知ることができる．出力する特徴語を文書中の単語や属性とすることで容易に分析の観点を変更することが可能である．

3.3　分析結果の活用

「ML 単位」の分析により，ML 間の関連を視覚化することができた．この結果から，ML の統廃合や，ML 参加者への関連 ML の紹介などを行うことができる．

特徴語として【送信者】を選択することにより，どのような人たちがどのようなトピック（知識）を持つかがわかる．これを社内の知識マップとして活用することも期待できる．

ML で様々なトピックがある場合は，ML 単位の分析では個々のトピックが他のトピックと干渉しあうが，月単位にすることにより多くの話題を含む状態を回避することができ，共通の話題をもつ ML 同士の関連が見やすくなる．

4．おわりに

本研究では，属性情報を等価的に扱うことができる情報検索システム"CoreExplorer"を利用することで，分析者の観点を自由に変更でき，分析対象を自在に絞りながら分析を行うシステムを作成した．

ML を対象に類似 ML を探し出す観点で分析を行ったところ，同様のトピックの存在や，クロスポストされているメールの存在から関連する ML 同士を見つけることができた．また，同様のトピックに関わる送信者同士の関連を把握できることが確かめられた．

ML 以外にも，システムを変更することなく様々な分析を行うことができる．例えば，保守日報から月ごとの不良発生を監視する分析や，年齢・性別・文書中の単語などで分析対象を絞りながら意見を抽出するアンケート分析などにも適用可能である．

本分析システムでは 2.1 節テキスト検索エンジンの(3)文書間の関連情報をそのまま利用しており，3.1 節の ML 単位，月単位の準備が必要になっている．これらの準備の代わりに，文書間の関連を複数文書をまとめた文書群同士の関連に拡張することで，ML 単位や月単位，スレッド単位など動的に文書の粒度を決定し関連トピックを判定することで関連トピック判定精度を向上できると思われる．

今後，分析に用いる観点を，特徴語から文を解釈した概念が扱えるように拡張し，概念に関わる ML・人・部署など分析を通してオントロジー構築等が行えるようにしていきたい．

参考文献

[1]塚原朋哉：情報の視覚的検索方法, FIT（情報科学技術フォーラム）2003, E-043, pp.179-180, 2003.