2012年6月15日　統計数理研究所　オープンハウス

テキストマイニングによる文章の分類と統計科学

高橋 久尚*　　サービス科学研究センター 特任研究員

テキストマイニングとは

- ◆テキストから情報を取り出す技術のこと
- ◆SNSなどの発達に伴い、消費動向調査、商品の評判を分析するツールとして注目されている。評判分析など。
- ◆著者の特定（源氏物語）など多様な用途でニーズはある

手法

- ◆文章を単語に分解（形態素解析）
  - ➢ 形態素解析作成時の辞書に依存、85%程度の精度
  - ➢ 隠れマルコフモデル（Viterbiアルゴリズム）
- ◆出現単語頻度を集計（出現単語頻度行列）
- ◆情報抽出
  - ➢ 判別分析、サポートベクターマシン、単純ベイズモデルなどの使用が考えられる
  - ➢ 出現単語情報のどの部分を使うのが最良か判断が必要

注意点

- ◆ビックデータがあれば、文章の分類も機械的にできる手法が見つかるのではと期待されるが、しかし、。。。
- ◆ビックデータであっても、カテゴリーに分けると意外に必要なデータは少ない。また、むしろデータの少ないところが重要なことすら多い。
- ◆データからの情報抽出に最良の方法を見つける技術（統計科学）が必要。

統計科学として

- ◆そもそも文章から得られる全情報を用いていない。単語出現頻度のみ。
- ◆言葉ネットワークのような係り受けなどを利用した、分析があるが精度の向上にはあまり有用ではない。多様性が大幅に増加してしまうのが原因か？
- ◆不完全情報による統計分析を完全情報にするための技術の開発が必要。
- ◆不完全データと聞けば、欠測値、逆問題などの用語が思い浮かぶが、文章を相手にしているので、別の手法が必要（？）

サービス科学として

- ◆アンケート調査の自由記述欄の分析など、ニーズはあるが社会には浸透していないのが実情か（？）
- ◆近年では、インターネット上の評判や何気なく書いてあるブログ、ツイッター情報を製品開発に生かすための道具とできないか、試みは多々ある。
- ◆コールセンターなどに寄せられる、製品の苦情相談の分析では使われているようである。
- ◆インターネット上の文章を分析することで、製品、サービスのニーズなどが分析できる基盤を作成できれば、新たな分析分野を確立できるはず。
- ◆新聞の文章の分析を用いた株価動向に関する研究はある。
- ◆データ・キュレーションの道具に成長させられないか、期待。

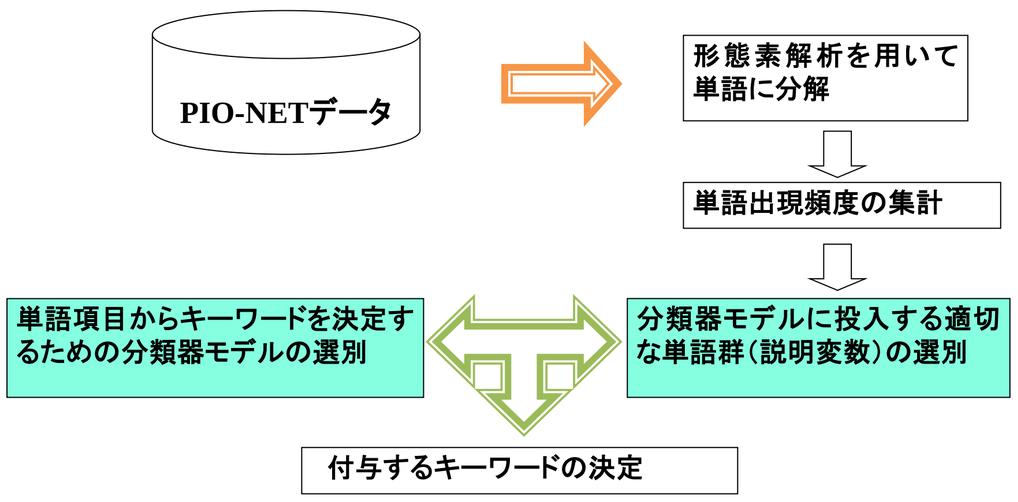
*産業技術総合研究所 協力研究員

PIO-NETデータの場合**

- 独立行政法人国民生活センターが運用。消費者苦情相談のデータ
- 文章に対して複数の分類コードを付与する。商品キーワード、相談内容キーワード
- 現在は相談員が付与。これを（半）自動化し、負担の軽減を目指す

手法

- 形態素解析は MeCab を使用
- 出現単語頻度を集計
- 判別分析、サポートベクターマシン、単純ベイズモデルを用いて
- 出現単語頻度を単純集計、χ^2 統計量、TF-IDF のそれぞれを用いて選別、それぞれの分類精度を比較。



分類精度は

χ2統計量を用いた単語選別による、線形判別またはSVMによる分類の精度が高い。単語数100程度のところで分類精度は最もよい。

「インターネット通販」の分析結果

		線形判別分析 案件数					SVM 案件数					単純ベイズ判別器 案件数				
	単語数	500	1000	2000	4000	8000	500	1000	2000	4000	8000	500	1000	2000	4000	8000
出現頻度による選別	50	0.856	0.868	0.890	0.871	0.889	0.872	0.897	0.927	0.889	0.915	0.470	0.527	0.567	0.541	0.650
	60	0.906	0.883	0.879	0.889	0.889	0.924	0.903	0.904	0.911	0.916	0.550	0.602	0.557	0.546	0.624
	80	0.900	0.889	0.879	0.892	0.894	0.904	0.918	0.899	0.912	0.914	0.548	0.607	0.584	0.552	0.625
	100	0.868	0.895	0.906	0.906	0.903	0.892	0.920	0.930	0.924	0.925	0.484	0.610	0.564	0.550	0.570
	200	0.848	0.877	0.909	0.916	0.911	0.884	0.905	0.921	0.925	0.926	0.520	0.553	0.550	0.556	0.552
	300	0.800	0.862	0.907	0.917	0.916	0.934	0.896	0.918	0.923	0.925	0.574	0.527	0.535	0.559	0.539
	400	0.700	0.856	0.897	0.919	0.917	0.886	0.907	0.920	0.924	0.924	0.556	0.534	0.513	0.569	0.544
	600	0.582	0.805	0.887	0.914	0.918	0.884	0.885	0.915	0.917	0.925	0.526	0.542	0.519	0.570	0.536
	800	0.674	0.698	0.859	0.911	0.922	0.860	0.883	0.908	0.916	0.924	0.498	0.524	0.521	0.555	0.541
	1000	0.602	0.506	0.819	0.902	0.923	0.796	0.876	0.899	0.912	0.922	0.502	0.502	0.536	0.552	0.548
χ^2 値による選別	50	0.854	0.877	0.908	0.904	0.901	0.866	0.901	0.931	0.931	0.929	0.518	0.491	0.472	0.609	0.640
	60	0.920	0.900	0.899	0.907	0.905	0.930	0.916	0.923	0.932	0.930	0.588	0.533	0.630	0.603	0.628
	80	0.892	0.916	0.902	0.915	0.909	0.910	0.919	0.918	0.935	0.930	0.564	0.588	0.601	0.564	0.569
	100	0.876	0.900	0.920	0.918	0.911	0.886	0.912	0.932	0.930	0.927	0.506	0.548	0.468	0.555	0.552
	200	0.838	0.877	0.917	0.926	0.918	0.882	0.894	0.919	0.930	0.929	0.534	0.522	0.515	0.540	0.556
	300	0.866	0.867	0.915	0.923	0.919	0.906	0.900	0.921	0.923	0.925	0.552	0.502	0.510	0.544	0.561
	400	0.674	0.870	0.911	0.921	0.920	0.868	0.903	0.918	0.923	0.923	0.540	0.561	0.497	0.537	0.563
	600	0.506	0.746	0.888	0.919	0.921	0.874	0.885	0.905	0.917	0.919	0.498	0.500	0.503	0.542	0.552
	800	0.618	0.673	0.856	0.917	0.924	0.878	0.877	0.902	0.909	0.919	0.542	0.504	0.498	0.527	0.549
	1000	0.580	0.597	0.821	0.906	0.924	0.816	0.870	0.890	0.910	0.919	0.502	0.500	0.502	0.516	0.538
TF-IDFによる選別	50	0.798	0.759	0.786	0.778	0.787	0.826	0.755	0.768	0.773	0.788					
	60	–	0.764	0.790	0.779	0.794	–	0.774	0.789	0.784	0.793					
	80	–	0.760	0.787	0.785	0.796	–	0.778	0.797	0.791	0.807					
	100	–	0.816	0.793	0.793	0.805	–	0.830	0.800	0.792	0.805					
	200	–	–	0.815	0.803	0.810	–	–	0.787	0.790	0.790					
	300	–	–	0.848	0.805	0.816	–	–	0.842	0.795	0.788					
	400	–	–	–	0.801	0.821	–	–	–	0.796	0.795					
	600	–	–	–	0.852	0.826	–	–	–	0.845	0.796					
	800	–	–	–	–	0.827	–	–	–	–	0.800					
	1000	–	–	–	–	0.828	–	–	–	–	0.821					

全体での分類精度

408分類コードの分類精度の平均は

該当	非該当	平均
87.40%	91.80%	89.6%□

特に悪いもの、良いものの例は

該当	最高値	サラダ油	99.0%
該当	最低値	契約	51.6%
非該当	最高値	コンサート	99.8%
非該当	最低値	契約	57.0%

**独立行政法人国民生活センターが独立行政法人産業技術総合研究所に委託した「PIO-NETの自然言語データを利用したデータ解析技術に関する研究業務」にかかる研究成果の一部