

# 多言語 Web テキストからの知識マイニングに関する研究

中川 裕志<sup>†</sup>, 二宮 崇<sup>†</sup> 吉田 稔<sup>†</sup> 清田 陽司<sup>†</sup> 佐藤 一誠<sup>††</sup>

<sup>†</sup> 東京大学情報基盤センター, n3@dl.itc.u-tokyo.ac.jp, {nimomi, mino, kiyota, issei}@r.dl.itc.u-tokyo.ac.jp

<sup>††</sup> 研究協力者: 東京大学 大学院 情報理工学系研究科, issei@r.dl.itc.u-tokyo.ac.jp

## 1. 概要

● Webテキストは知識の宝庫だが、大きすぎて探すのが難しい。Googleのようなキーワードによる検索エンジンでも思ったような結果が得られないことはままある。

● その問題を解決する有力な方法として文書をトピック毎に分類する研究が90年代から続けられた。90年代後半に86%程度の精度を実現したところで頭打ちとなった。このときは、下の図のように、「サッカー」「オリンピック」など単一のトピックに沿って分類する研究が行われた。

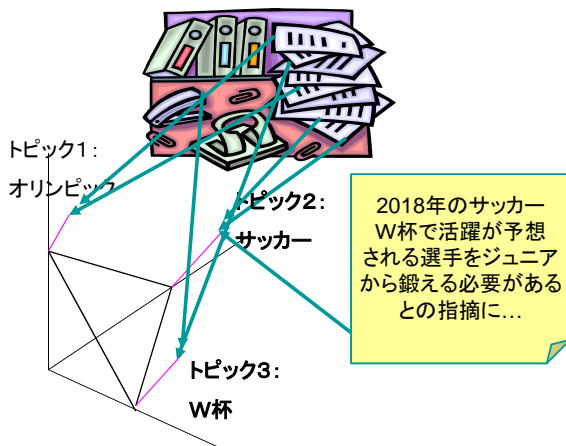


図1 単一トピックの文書分類  
Fig 1. Single Topic Text Classification

● 最近のWeb2.0の流れの中で作られているテキストは、wikipediaやブログに代表されるように、一つのテキストが複数のトピックを持つことが多い。例えば、次の文書：Wについて考えよう。

W杯といえばなんといってもサッカーが一番盛り上がる。アジア地区の予選でさえ視聴率は高い。そういえばオリンピックの予選もすごかった

文書：W

● この文書には「サッカー」「W杯」「オリンピック」「アジア」など多くのトピックが入っている。したがって、この文書を分類するにあたっては、それぞれのトピックに分類されていなければならない。

● しかし、少しでも関連するトピックに分類すると、

トピック毎に分類された文書数が大きくなりすぎて、せっかくの分類が役に立たなくなってしまいかねない。そこで、以下に述べるアプローチを採る。

## 2. 手法・アプローチ

● 与えられたトピックが各単語のどのような重み付きの分布で表わされるかを教師あり機械学習する。たとえば、「サッカー」というトピックは「W杯, 予選, アジア, ...」などの単語の学習された重み付き分布で表現される。

● 新たなテキストがやってきたときの分類は、こうして学習されたトピックがどのような配分比で混合しているテキストかを推論する。(下図を参照) 配分比のモデル化には高次元空間の扱いに適したディリクレ分布を用いる。

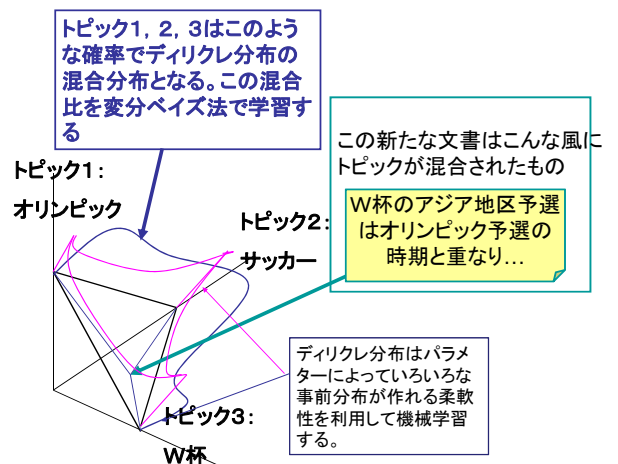


図2 多重トピックのテキスト分類  
Fig. 2 Multi-topic Text Clustering

## 3. 議論・考察

● この推論のために我々は効率の高い変分ベイズ法を工夫した。成果の一部を下に記載する。

### 成果 (一部)

- [1] 佐藤一誠, 中川裕志. Dirichlet Process Unigram Mixture Model に対する Collapsed 変分ベイズ法の適用, 情報処理学会論文誌, Vol.48 TOM19, pp.1-10, (2007)
- [2] Issei Sato, Hiroshi Nakagawa. Knowledge Discovery of Multiple-topic Document using Parametric Mixture Model with Dirichlet Prior, Thirteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.590-598, (2007)