1A1-03

知識検索サイトにおける有害情報のフィルタリング知識の表出化

Display Knowledge Filtering Harmful Information on Knowledge Searching Website

小林大祐[*1] 松村真宏[*2] 石塚満[*1]
Daisuke Kobayashi Naohiro Matsumura Mitsuru Ishizuka

[*1] 東京大学大学院情報理工学系研究科 [*2] 大阪大学大学院経済学研究科
The University of Tokyo Osaka University

It is important for knowledge searching website such that one user asks and another user answers to reduce harmful question and answer for improving quality. Now, harmful posts are filtered by human, so costs for filtering with increasing posts and different standard due to filtering by many people become problem. This paper tries to classify such harmful posts automatically, and display classification knowledge from labeled posts and hold the knowledge in common.

1. はじめに

近年、Web の急速な普及により Web 上に膨大な量の情報が蓄積されるようになってきた。また、それらの膨大な情報が集まっているサイトが存在している。

そのようなサイトの一つである、ユーザからの質問にユーザが答える知識検索サイトでは、サイトの質を高めるために不適切な質問や回答を減らすことが重要である。現在そのような不適切な投稿のフィルタリングは人手で行われているが、投稿数の増加に伴うフィルタリング労力のコスト、およびフィルタリングの基準が曖昧で人によって異なることが問題となっている。

本研究では人手でフィルタリングされた投稿から、フィルタリングする際に暗黙的に用いられる分類知識を表出化し、フィルタリングの自動化と分類知識の共有を試みる。

本稿では、2 章では Web フィルタリングに関する関連研究について述べる。3 章では本研究の手法、4 章では実験結果について述べ、5 章でまとめる。

2. 関連研究

Web において有害な情報をフィルタリングする研究は盛んに行われている。ここではその中から 2 つの研究を紹介する。

2.1 スパムメールフィルタへのテキスト分類の適用

[Draucker 99]では、テキスト分類の手法を用いてスパムメールフィルタの分類を行っている。分類の元となるタイトルと文章のどちらか片方もしくは両方、特徴値として TF を用いた SVM や Rocchio のアルゴリズム[Schapire 98] [Joachims 97]、boosting algorithm[Freund 96]などを用い、様々な手法を比較している。

なお、この中で精度がよかったのは、素性に binary 値を用いた場合の SVM と boosting algorithm で、もととなる文章は、メールのタイトルと本文をともに使った時であった。一番精度がいい場合には、精度は 98%ほどとなった。

2.2 有害な Web テキスト文のフィルタリング

[Grilheres 04]では、Web 上の有害な文章のフィルタリングを行った。しかし、一口に有害な文章といっても、様々な種類、内容を持った文章が存在する。この研究では、有害かそうでないかラベル付けされたサイトのデータセットを用いて、それを分類している。

例えば、麻薬に関する記述がされているサイトがあるとする。その内容が「麻薬の売買」に関するサイトであるならば、そのサイトは明らかに有害であると考えられる。しかし、「麻薬への注意を喚起する」ような内容のサイトであるならば、そのサイトは有害であるとはいえないであろう。

このように単純に bag-of-words で分類してしまうと、フィルタリングすべきでないサイトがフィルタリングされる可能性がある。

この研究では、SVM による分類、もともと危険なアドレスだと分かっているアドレスのパターンマッチングやサイトに存在している画像の色の割合それぞれを用いて分類を行い、それぞれの出力結果をさらに SVM で分類したりすることで、90%ほどの精度を得ている。

3. 提案手法

本稿では、ヤフー株式会社が提供している Yahoo!知恵袋 (http://chiebukuro.yahoo.co.jp)について、禁止行為に該当する情報を用いて分類を行う。Yahoo!知恵袋は、あるユーザが質問内容をあらかじめいくつかに分けられたカテゴリに投稿し、別のユーザがそれに対する回答を投稿するという知識検索サイトである。また、質問･回答を閲覧したユーザの投票で質問に対する回答の中から 1 つをベストアンサーとして選ぶことができる。

3.1 禁止行為

このサイトには、利用する際のガイドラインがある。その中に禁止行為というものが示されており、それらに抵触する投稿は削除されるとされている[*1]。このガイドラインには、以下のような投稿が禁止行為として示されている。

- いやがらせ、悪口、脅し、あるいは有害な内容の掲載など、他人を攻撃したり、傷つけたりする目的で利用すること
- わいせつな内容や不愉快なデータを公開すること
- 商業目的や広告目的で利用すること
- 質問と関係のないことを書き込むこと
- Yahoo! JAPAN が予定していない目的で本サービスを利用すること
- 著作権者の許可を受けずに著作物を公開するなど、第三者の知的財産権を侵害したり、侵害を助長すること
- プライバシー侵害の恐れがある事実やデータを公開すること
- その他、Yahoo! JAPAN が不適切だと判断する行為

連絡先：小林大祐, 東京大学情報理工学系研究科, 東京都文京区本郷 7-3-1, d-koba@mi.ci.i.u-tokyo.ac.jp

[*1] http://chiebukuro.yahoo.co.jp/docs/guidelines.html

表 1 禁止行為に抵触する投稿の例

- 風、結構強くないですか？
- 何ラーメンが一番、うまいかなぁ？
- うわ･･･今の服なんだ！？化粧も濃いし･･･靴のヒールも高いし･･･勘違いもいいところですよね？
- 他人の持ってる大事な物を何でも欲しがる超ワガママなくそがきジーがホチ～～～って言って駄々こねてるんだって！誰かお灸をすえてやるやつはおらんのか！！！

このような禁止行為に抵触する投稿の例を表 1 に載せる。

この中で、「いやがらせや悪口、脅しの掲載」などについては bag-of-words を用いたテキスト分類を行えばよい。また、同じ文章の投稿をいろいろなカテゴリに対して行ういわゆる「マルチポスト」と呼ばれる投稿に関しては各投稿のハッシュをとるなどすればよいし、また、人によってフィルタリングの基準が異なるということもないであろう。

しかし、実際の削除データの中で、もっとも禁止行為に当てはまるとされた件数が多いのは、「Yahoo! JAPAN が予定していない目的で本サービスを利用すること」に関する投稿である。サイトが想定していないような文章などを質問として投稿することによりサイトの質が下がってしまうため、削除が必要である。投稿に使われている単語そのものはサイトで予定しているような質問と変わらないため、自動的に削除対象かどうか判断するのは難しい。

3.2 禁止行為に抵触する投稿の分類

このような投稿を分類するために、文章をグラフ化する事を考える。この種の投稿では、単語間の共起頻度が低いという事が見受けられたので、このような手法をとることとした。これを正しい文章と共起頻度を比較することによって、その重なり具合が低い場合には、禁止行為に抵触する投稿として分類することとする。

3.3 データセット

今回は、Yahoo!知恵袋の禁止行為に抵触するデータを削除された例のデータセットとして用い、削除されなかった例のデータセットとしては、Yahoo!知恵袋のベストアンサーを用いることとする。

また、正しい文章の語の共起頻度としては、日本語版 Wikipedia (http://jp.wikipedia.org)の文章から共起頻度を取ることとした。Wikipedia は、ユーザが単語とその意味を登録することができるサイトで、全世界の言語について語とその意味が掲載されている。日本語版では 19 万語もの語が登録されている。

Wikipedia のデータセットを用いる理由は、新聞コーパスなどを用いると、最新の単語などが入っていないなどの原因で Web 上の文章を正しく分類できない可能性があると考えられるのに対して、Wikipedia のデータセットであれば最新の単語などが登録されている可能性が高く、投稿データに最新の単語が出現した場合に、見逃す可能性が低いと考えたためである。

4. 実験

3.3 節のデータセットを用い、正しい投稿の共起頻度を取得する。まず、全ての Wikipedia のページから、名詞による N-Gram を抽出し、ある一定の出現回数(tf)、出現文章数(df)を持つものを抽出する。全ての Wikipedia に登録された単語を形態素解析し、名詞のみで N-Gram(N≦5)、出現回数(tf)が 10 回以上で出現文章数(df)が 2 回以上のものを抽出したところ、12277 の語が抽出された。

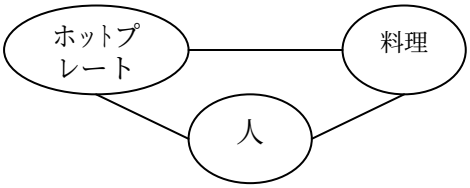

図 1 「ホットプレートを使う料理で、7 から 8 人くらいで食べられる料理教えて下さい。」という投稿文から得られる共起グラフ

これらの単語間の共起頻度を Wikipedia を用いて取得する。

また、実際に Yahoo!知恵袋内において「いたずら投稿」として削除されたデータ、またベストアンサーに選ばれた回答、役に立つ質問として掲載されている質問から共起グラフを作り Wikipedia から得られた共起頻度と比較し、実際にどうなったかを報告する。

5. まとめ

本研究では、禁止行為に抵触する投稿の中でも、「予定していない目的でのサービスの利用」にあてはまる投稿の分類を行い、それを行うことができる事を確認した。今後は、それ以外の投稿の分類なども実装して、実際の知識検索サイト中で使用できるような分類器の作成を目指したい。

謝辞

本研究では、ヤフー株式会社様より、Yahoo!知恵袋のデータをご提供をいただき、分析に利用させていただきました。ご協力いただきましたヤフー株式会社様には記してお礼を申し上げます。

参考文献

[Drucker 99] Support Vector Machines for Spam Categorization, H. Drucker, eith C. Wu and V. Vapnik. IEEE Trans. On Neural Networks, vol. 10, number 5, pp. 1048-1054. 1999.

[Schapire 98] R. E. Schapire, Y. Singer, and A. Singhal, “Boosting and Rocchio applied to text filtering,” in Proc. 21st Annu. Int. Conf. Inform. Retrival, SIGIR, 1998.

[Joachims 97] T. Joachims, “A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization,” in Proc. 14th Int.Conf. Machine Learning, D. Fisher, Ed. San Mateo, CA: Morgan Kaufmann, 1997.

[Freund 96] Y. Freund and R. E. Schapire, “Experiments with a new boosting algorithm,” in Machine Learning: Proc. 13th Int. Conf. San Mateo, CA: Morgan Kaufmann, 1966, pp. 148-156

[Grilheres 04] B. Grailheres, S. Brunessaux, P. Leray, Combining Classifiers for harmful document filtering, RIAO’2004, Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval, 2004.