ОТЗЫВ
на магистерскую диссертацию
«Классификация текстов электронных писем для упрощения поиска»
Чжао Шугуан

Диссертация посвящена актуальной теме автоматической категоризации текстовых документов (электронных писем). Каждому письму система категоризации должна назначить ключевое слово (тег), по которому пользователь сможет составить представление о содержании документа. Все теги должны быть представлены в виде облака тегов, анализируя которое пользователь может выбрать тег, отвечающий его текущим информационным потребностям, и далее по выбранному тегу получить доступ к соответствующему множеству писем.

В диссертации рассмотрены следующие основные вопросы:

1. Разработана архитектура системы, в которой Веб-сервер обеспечивает взаимодействие с клиентом (загрузка писем, отображение облака тегов, выдача результатов), а сервер баз данных выполняет как функцию хранения данных, так и их обработки (в том числе и весьма сложной – разбор текстов, кластеризация писем, выделение ключевых слов/тегов).
2. Обработка данных. Как уже говорилось выше, обработка данных весьма сложна и включает в себя такие шаги как выделение слов из документа (и их фильтрацию), преобразование неправильных глаголов к инфинитиву (система ориентирована на английский язык), удаление стоп-слов, удаление некоторых окончаний (вариант стемминга), сбор статистики (подсчет частот слов), выделение ключевых слов по критерию TF*IDF (вес слова).
3. Кластеризация писем. Здесь применяется простой алгоритм кластеризации, основанный на вычислении меры подобия между письмами, которая задается косинусом угла между векторами, состоящими из весов входящих в письма ключевых слов. В частности говорится, что документы, мера подобия между которыми больше средней величины меры подобия по всем парам писем, мера подобия между которыми строго больше 0 и строго меньше 1, относятся к одному кластеру.
4. Выделение тегов – ключевых слов имеющих максимальную среднюю частоту в документах данного кластера.

Нужно отметить следующие недостатки работы:

1. Пример, приведенный в тексте диссертации, слишком мал (только 7 писем) для оценки качества предложенных алгоритмов.
2. По просьбе рецензента автор проверил работу алгоритма для большего числа писем (32), и тут вскрылись некоторые проблемы. В частности, реализация алгоритма не соответствует описанию алгоритма, приведенному в тексте диссертации. Это приводит к тому, что построенные кластеры имеют значительно меньшие размеры и их число велико. Например, для 32 писем было построено 14 кластеров, и при этом два кластера получили один и тот же тег.

В целом можно отметить, что решаемая задача является весьма важной и интересной, подход к решению задачи, основанный на кластеризации писем, является правильным. Однако используемый метод кластеризации (и особенно ошибки в его реализации) привели к тому, что система еще не может использоваться на практике и требует значительной доработки. В связи с вышеизложенным считаю, что Чжао Шугуан достоин присуждения ему степени магистра, но оцениваю его диссертацию «Классификация текстов электронных писем для упрощения поиска» на оценку «удовлетворительно».

Рецензент,
к.ф.-м.н., доцент  В.Ю. Добрынин

23.06.11