CEO Russia Гиргидов Рубен

Автоматизированные методы определения эмоций и отношения потребителя к продукту.

AlephOne

Positive – Negative Category Analysis

Задача.

- Основа работы системы это категоризация текстов на базе различных критериев.
  - Базовый алгоритм предполагает произвольное количество категорий.
  - Критерии и их веса определяются как нечеткие логические конструкции и выводы.
  - При анализе текстов использовать не только «лингвистические» критерии, но и сопутствующую информацию
  - Алгоритм предполагает обучение с минимальным участием человека

Категории и начальные требования

- Адаптация алгоритма категоризации
  - Алгоритм обучаемый с учителем
  - 2 Категории:
    - Positive
    - Negative
  - Тренировочный корпус текстов ~ 50% заведомо позитивных сообщений и 50% отрицательных:
    - 100 сообщений с сайта Buy.com форум телефонов
    - 100 сообщений с сайта Amazon форум бытовой техники
  - Словарь термов должен выделяться автоматически

Шаги обучения системы 1

- Генерация словаря термов
  - Генерация словаря на базе стандартных словарей английского языка
  - Генерация словаря на базе Wiki
  - Генерация словаря на основе базы сообщений из корпуса
- Генерация правил разбора сообщения
  - Определение важнейших параметров сообщения для каждого форума:
    - Частотные
    - Текстовые
    - Мета информационные

Шаги обучения системы 2

Частотные критерии	Текстовые критерии	Мета информационные критерии
Частота появления терма в тексте сообщения	Принадлежность терма к названию форума сообщения	Дата сообщения
Совместность появления различных термов.	Принадлежность терма к теме сообщении	Количество сообщений в ветке
Частота терма в целом по форуму	Принадлежность терма к тексту сообщения	Форматирование текста
…	…	…

Betria systems

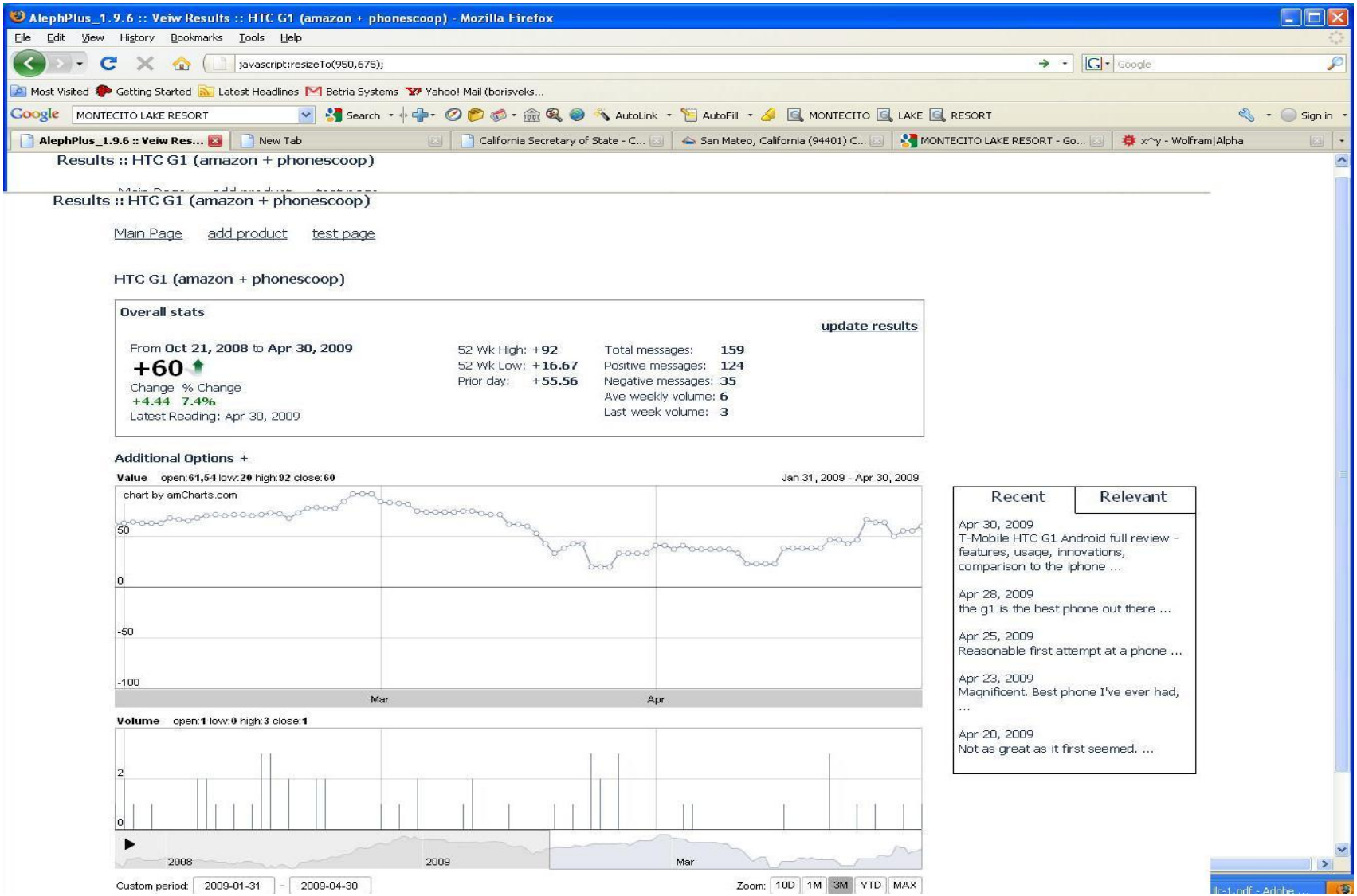
Шаги обучения системы 3

- Составление функционала принадлежности к категории
  - Нечеткие логические конструкции
  - Весовые коэффициенты
- Подстановка корпуса положительных и отрицательных сообщений для определения весовых коэффициентов
  - 100 сообщений с сайта Buy.com форум телефонов
  - 100 сообщений с сайта Amazon форум бытовой техники и электроники

BlackBerry Storm

Betria systems

Android G1



Betria systems

iPhone3G

Betria systems

Результаты исследования 1

- Словарь термов
  - Сгенерированный словарь практически не повлиял на точность отнесения того или иного сообщения к категории, но повлиял на уверенность отнесения отдельного сообщения к категории (чем обширней словарь, тем хуже результат)
  - Худший результат у формального словаря английского языка (результаты не валидны)

  Вывод: метод определения термов, использованный в нашей компании в целом оказался эффективен для английского языка

Результаты исследования 2

- Словарь Stopword отклонение составило не более 5-10%
  - Google stopwords средний результат
  - Wordnet stopwords худший результат
  - Созданный нами лучший результат

Вывод: вероятно сказалась привязка фильтра к «форумному сленгу». У Google средний результат говорит, что они вынуждено идут на компромиссы, т.к. имеют дело с текстами всех тематик одновременно. У Wordnet худший результат, т.к. они имеют дело с текстами больших объемов и достаточно чистыми.

Результаты исследования 3

- Тематика обучающего корпуса текстов практически не имеет значения. Ее можно определить как техническая.
  - Машины,
  - Телефоны
  - Бытовая техника
  - Электроника
- Необходимо соблюдать баланс между положительными и отрицательными сообщениями (не более 20%)

Вывод: Точность определения составила 75%±10% вне зависимости от обучающей выборки.

Результаты исследования 4

- Наибольшую сложность представляла величина уверенности отнесения сообщения негативным или позитивным текстам. Увеличение Длины сообщения только ухудшало ситуацию.
- Есть некоторые темы, для которых не существует позитивных сообщений. К ним относятся:
  - Политика
  - Бюрократические процедуры
- Алгоритм оказался достаточно устойчив к сообщениям с условиями (например: «вроде бы не плох»)

В целом использование методов категоризации текстов применительно к эмоциональным категориям применимо, но осложняется, требованием единственности эмоции на текст.

Вывод

Использование методов категоризации текстов, применительно к эмоциональным категориям работает, но осложняется, требованием единственности эмоции на текст.

Текущее состояние

В настоящее время исследования приостановлены и разработка продукта заморожена, в связи с отсутствием коммерческого спроса

Вопросы

Конец

Автоматизированные методы определения эмоций и отношения потребителя к продукту.

Рубен Гиргидов

ruben@betria.com

Betria systems