ВЫЯВЛЕНИЕ РЕЧЕВЫХ КОНСТРУКЦИЙ ПОВЫШАЮЩИХ ТОЧНОСТЬ РАБОТЫ СИСТЕМЫ ИНФОРМАЦИОННОГО ПОИСКА

IDENTIFICATION SPEECH DESIGN IMPROVES THE ACCURACY OF THE INFORMATION SEARCH SYSTEM

D. Potapov
D. Akimov

Summary. The potency of information search system is measured by completeness and accuracy, while improving the performance of one of these word processing algorithms on the other indicator is the opposite effect, so that the two quantities are inversely related. To improve the overall completeness and accuracy of information should be used in providing a number of algorithms the amount of increase in both values, due to the fact that, in particular when they are used for inverse index smoothed decline in favor of the indicator is to improve the algorithm aims.

And this article is an algorithm for identifying the most informative elements of the text, which can give higher weight in comparison with other elements.

The relevance of the study due to the list of priority directions of development of science, technology and engineering of the Russian Federation, approved by Presidential Decree of July 7, 2011 № 899.

Keywords: data Library; linguistic analysis; indexing.

Потапов Дмитрий Анатольевич
Аспирант, Московский технологический университет
potdmiana@mail.ru
Акимов Дмитрий Александрович
К.т.н., Московский технологический университет
Press@smartinterface.ru

Аннотация. Эффективность работы системы информационного поиска измеряется полнотой и точностью, при улучшении одного из этих показателей алгоритмами обработки текста на другой показатель оказывается обратное воздействие, таким образом обе величины являются обратно зависимыми. Для общего повышения полноты и точности информации следует применять ряд алгоритмов оказывающих в сумме повышение обеих величин, за счёт того что в частности при их применении сглаживается снижение обратнозависимого показателя в угоду показателю на улучшение которого нацелен алгоритм.

А данной статье будет предложен алгоритм выявляющий наиболее информативные элементы текста, которым можно будет раздать более высокие веса в сравнении с другими элементами.

Актуальность исследования обусловлена перечнем приоритетных направлений развития науки, технологий и техники Российской Федерации утвержден Указом Президента Российской Федерации от 7 июля 2011 г. № 899.

Ключевые слова: библиотеки данных; лингвистический анализ; индексирование

Проведение анализа мнений пользователей сети интернет в первую следует начать лингвистического анализ текстов сообщений. Результативность лингвистического анализа измеряется по полноте и точности классификации сообщения. На вход могут поступать сообщения разной размерности и тематик, на выходе будет получена библиотека размеченных сообщений пользователей с указанием принадлежности к тематикам и датой актуальности.

Сообщения пользователей слишком разнородны что бы можно было прийти к единому методу проведения лингвистического анализа. Опираясь на эмпирические знания выделим диапазон размерности текста, который достаточен что бы нести в себе смысловую нагрузку и мнение, но в тоже время не является слишком большим многосмысловым текстом, например статьёй. Размер оригинального текста сообщения в нашем случае должен быть в диапазоне от 10 до 100 слов.

С течением времени появляются новые темы обсуждения, поэтому использовать только заранее размеченную библиотеку не рационально, т.к. она либо потеряет актуальность либо будет требовать периодической ручной актуализации. Однако часть тематик возможно предусмотреть заранее, например такие темы как: революция, праздник, ЧС, правительство. Заранее размеченная библиотека позволит использовать более жёсткие правила на автоматическое пополнение библиотеки новыми речевыми конструкциями. В библиотеку должны попадать речевые конструкции размером не более 3 слов, в случае если конструкция является единым понятием, то система должна производить разделение конструкции на несколько элементов. Так например элемент «Коммунистическая партия Российской Федерации» может быть преобразована в элементы «коммунисты» и «Россия», в таком случае текст описывающий события связанное с КПРФ должен ссылаться на оба понятия и «коммунисты» и «Россия».

Использование в библиотеке слов, а не фраз обусловлено тем фактом, что фразы обладают худшими статистическими характеристиками чем одиночные слова [1]. Фразы допустимо использовать только как идентификаторы конкретных людей или гео-объектов. Так как библиотека будет представлять из себя только набор

терминов, т.е. хранить только значение, то её реализация будет в виде пополняемого списка.

Так как сообщения пользователей короткие, искать термины для библиотеки тематик следует только массивах текста агрегированных из некоторого количества сообщений. С целью повышения качества сырья в единый массив следует объединять сообщения смежных направлений. Исходя из вышесказанного сформируем требования к алгоритму поиска терминов для библиотеки тематик:

1. Выявление смежности сообщений
2. Взвешивание слов
3. Определение диапазонов весов, достаточных для внесения слова в список тематик

Смежными предложениями будут считаться предложения имеющие наибольшую силу связи. Сила связи вычисляется по формуле:

$$F(s_{n,i}) = \sum_{n=1,i=1} Q(s_n) * Q(s_i) \quad [3]$$

Формула расчёта силы связи предложений

Где:

$Q(S_n)$ — количество вхождений конкретного слова s в предложение S_n

$Q(S_i)$ — количество вхождений конкретного слова s в предложение S_i

По представленной формуле будет вычисляться коэффициент силы связи двух предложений. Поученные значения удобно хранить в виде квадратной матрицы, в которой строки и столбцы являются симметрично упорядоченными сообщениями. Все элементы матрицы являются неотрицательными целыми числами, а главная диагональ приведена к нулю. Сама матрица всегда будет симметричной при правильном упорядочивании столбцов и строк.

Взвешивание слов будем производить по формуле вероятности встречи слова в конкретном массиве текста:

$$p = {}^{N_i}/_{N_{max}}$$

Формула расчёта вероятности

Где:
N_i — количество повторений слова *i* в массиве
N_{max} — количество всех слов в массиве

Опираясь на законы Ципфа и исследования, которые показывают что наиболее значимые для текста слова лежат в средней части графика [2], добавим условие, что для пополнения библиотеки будет использоваться только малая часть слов, расположенная на графике максимально близко к нулю. Кол-во таких слов должно быть зависимо от размера анализируемого массива, для построения зависимости будем использовать двоичный логарифм.

Что бы выявить наиболее важные для определения смысла сообщения слова воспользуемся дистрибутивно-семантическим моделью vector space model (модель векторного пространства). Принадлежность сообщения к тематике будем выявлять через меру TF-IDF. На вход будут подаваться коллекции сообщений, но в отличии от общепринятого расчёта индекса для всех слов, будет производиться расчёт индекса только для тех слов, которые присутствуют в библиотеке тематик.

Воспользуемся немодифицированной формулой tf*idf.

TF (Term frequency) — рассчитывается как количество вхождений конкретного термина в конкретное сообщение, делённое на общее количество слов в этом сообщении.

$$tf(t,d) = \frac{n_i}{\sum_k n_k}$$

Формула расчёта TF

IDF (Inverse document frequency) — рассчитывается как инверсия частоты, с которой некоторое слово встречается в сообщениях коллекции.

$$idf(t,D) = \log \frac{|D|}{|d_i \supset t_i|}$$

Формула расчёта IDF

Где
t — термин
d — (data set) сообщение
D — количество сообщений
n_i — количество термина *t* в тексте сообщения
n_k — количество терминов в тексте сообщения
$|d_i \supset t_i|$ — количество сообщений, в которых встречается t_i, когда $n_i \neq 0$

Мера TF-IDF является произведением двух сомножителей:

$$\mathrm{tf-idf}(\mathrm{t},\mathrm{d},\mathrm{D}) = tf(t,d) \times idf(t,D)$$

Таким образом большой вес в TF-IDF получат слова с высокой частотой в пределах конкретного документа и с низкой частотой употреблений в других документах.

Распределение весов по представленной формуле позволит оценить значимость слова только с точки зрения вхождения в документ, порядок следования и синтаксические роли слов не учитываются.

Модель TF*IDF применяется в общей практике для объёмных текстов, для проверки эффективности применения модели в отношении сообщений пользователей социальной сети необходимо провести эксперимент.

Для оценки результативности выявления ключевых слов был разработан модуль, реализующий данный алгоритм на языке python.

```
def index_tf (text):
    tf_result = collections.Counter(text)
    for i in tf_result:
        tf_result[i] = tf_result[i]/float(len(text))
    return tf_result
```

Алгоритм расчёта меры tf

```
def index_idf(word, corpus):
    return len(corpus)/sum([1.0 for i in corpus
    if word in i])
```

Алгоритм расчёта меры idf

В качестве входных данных использовались тексты сообщений пользователей, с нормализованными формами слов. На выходе для каждого сообщения был получен набор терминов из библиотеки тематик встретившихся в тексте сообщения, и веса важности терминов в рамках сообщения.

Сформулированный подход использования библиотеки готовых речевых конструкций для поиска затрагиваемых в сообщении тем позволит разработать математическую модель выявления актуальных трендов, основанных на мнениях пользователей сети интернет.

ЛИТЕРАТУРА

1. Lewis D.D., An evaluation of phrasal and clustered representations on a text categorization task. In Proceedings of SIGIR92, 15th ACM International Conference on Researchand Development in Information Retrieval (Kobenhavn, DK, 1992), pp. 37–50., 1992
2. Kechedzhy K.E., Usatenko O. V., Yampol'skii V. A. Rank distributions of words in additive many-step Markov chains and the Zipf law (англ.) // Phys. Rev. E..— 2004.
3. Яцко В.А., Имметричное взвешивание терминов, Символ науки, (2015), 12–1 (декабрь), 87–90.

Журнал «Современная наука: актуальные проблемы теории и практики»