Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества

Libraries and Associations in the Transient World: New Technologies and New Forms of Cooperation

Материалы конференции
Conference Program

Том 1
Volume 1

г. Евпатория, Республика Крым, Украина
10-18 июня 1995 г.
Eupatory, Republic of Crimea, Ukraine
10-18 June 1995

N–граммные методы обработки текстовой информации

N–gram Techniques for Digital Information Processing

Мазов Н.А.
Объединенный институт геологии, геофизики и минералогии Сибирского отделения РАН, Новосибирск, Россия

Mazov N.A.
Joint Institute of Geology, Geophysics and Mineralogy, Siberian branch of Russian Academy of Sciences, Novosibirsk, Russia

A recognition technique based on utilization of N-grams, portions of natural language words, is considered for computerized classification and detection of spelling mistakes.

Развитие современной науки, техники и общества вызвало к жизни такой большой поток научно-технической и социальной информации, что традиционных методов ее накопления, систематизации и переработки становится явно недостаточно. Эффективным решением этой важной проблемы является автоматизация процессов обработки больших массивов текстовой информации, что предполагает более широкое использование автоматических процедур обработки текстов. Это позволит значительно повысить качество научно-исследовательских и экспериментальных работ и сократить сроки их выполнения.

Такой подход, в свою очередь, требует модификации классических процедур отбора информации (например тех, в которых запрос формулируется с использованием терминов словаря системы и булевой логики), разработки ряда новых, более адекватных природе и приемам обработки текстовой информации (например, автоматическая классификация массива документов на основе использования статистической теории распознавания образов).

В настоящем докладе рассматривается один из методов автоматического распознавания, который основан на использовании N-грамм слов естественного языка текстов документов применительно к задачам автоматической классификации и обнаружения орфографических ошибок в текстах документов.

1. Автоматическая классификация, формирование массива документов на основе отбора документов из различных баз данных.

В информационной практике широко известен классический подход к отбору информации, в котором формулируется поисковый запрос с использованием терминов, отражающих тематику информационной потребности. Имеется широкий спектр информационно-поисковых систем, реализующих этот подход. В большинстве таких систем пользователь имеет возможность уточнить информационное значение терминов при помощи булевой логики или присвоения им весовых коэффициентов. Однако формулирование поисковых запросов является непростой задачей для неподготовленного пользователя, поскольку такой формальный подход к формулированию запроса принуждает его приспосабливаться к системе, что создает для него свои неудобства. В связи с этим требуется наличие информационных работников, к профессиональному уровню которых предъявляются высокие требования.

Исходя из этого предлагается альтернативный подход, исходной посылкой которого является тот факт, что пользователь перед поиском обладает информацией, соответствующей его информационной потребности. На практике известны методы отбора документов, базирующихся на использовании статистических характеристик текстов. В качестве таких характеристик могут выступать частота появления терминов в базе данных и документах пользователя, совместная встречаемость терминов в документах и др.

Однако большинство таких методов предполагает предварительную канонизацию текстов документов, что прямо или косвенно связано с процедурами морфологического анализа или процедурами выделения информативных элементов терминов документов, где под информативным элементом текста документа понимаются такие элементы, которые однозначно передают смысловое содержание данного элемента.

К недостаткам этих методов следует отнести трудоемкость и длительность неформализованного этапа подготовки списков терминов с вытекающими проблемами по их организации и ведению.

В связи с вышеизложенным заслуживают внимания статистические методы выделения информативных элементов текстов документов и автоматической классификации, использующие в своей работе частотные характеристики различных фрагментов слов текстов документов. Одно из основных преимуществ такого подхода состоит в том, что он применим для любого естественного языка и не требует участия специалистов в конкретном языке. Вторым немаловажным фактором является то, что при использовании фрагментов постоянной длины возможно установить максимальное количество фрагментов (N-грамм) для данного языка.

Так, например, для русского языка при алфавите в 30 букв количество N-грамм при N=3 равно 27000. Однако, как показывает практика, в реальных текстах реализуется не более 25-30 процентов N-грамм от общего допустимого их числа, т.е. для русского языка их не более 7000.

Итак, метод основан на использовании вероятности появления цепочки букв N-го порядка (N-грамм) в анализируемых текстах.

В качестве классификатора для анализируемого множества документов применялась функция правдоподобия, а значимость тех или иных N-грамм определялась по критерию Хи-квадрат.

Существенной отличительной чертой предлагаемого метода использования N-грамм в отличие от известных полиграммных методов является то, что при их определении учитывается и позиционность N-грамм в словах текстов документов. Этот факт позволяет существенно повысить распознавательскую способность предложенного метода.

По предложенному методу был проведен эксперимент классификации (отбора) документов на отечественной документальной базе данных ВИНИТИ "Коррозия" общим объемом около 10 тысяч документов. В документах обрабатывались поля: заглавие, ключевые слова и рефераты. В качестве N-грамм использовались триграммы. В качестве экзаменационной выборки использовался массив документов базы данных "Химия" объемом более 30 тысяч документов. В результате анализа экспериментальных данных были получены: полнота — 90 процентов и точность 70 процентов.

Экспериментальные данные, полученные автором, показывают, что отбор (классификация) документов с использованием предложенного метода является хорошей альтернативой использованию традиционных методов отбора (поиска) при формулировании поисковых запросов с помощью булевой логики.

2. Автоматическое обнаружение орфографических ошибок в текстах документов

Одним из наиболее трудоемких и дорогостоящих процессов обработки информации является процесс обнаружения орфографических ошибок. Наряду с неуклонным ростом объемов научно-технической информации вводимых в ЭВМ возрастает и острота этой проблемы. Поэтому в последние годы увеличился интерес к данной проблеме, и в настоящее время разработано множество различных методов как у нас в стране, так и за рубежом, которые позволяют частично или полностью решать эту проблему.

Анализ работ автором по автоматическому обнаружению орфографических ошибок приводит к выводу о наличии сегодня следующих подходов:

— N-граммный, основанный на использовании допустимых буквенных сочетаний; — словарный, основанный на использовании эталонных и частотных словарей. Обнаружение орфографических ошибок в словарных системах бази

руется на пословном контроле текстов документов. Выделенные словоформы канонизируются процедурами морфологического анализа и затем эти словоформы сопоставляются со словоформами эталонного словаря. После этой процедуры пользователю выдается список правильных и ошибочных слов.

Однако, несмотря на все преимущества словарных систем обнаружения ошибок, они громоздки ввиду того, что они так или иначе используют в своей работе огромные эталонные словари. Это, в свою очередь, делает проблематичным построение гибких машинонезависимых систем.

Поэтому, на наш взгляд, несмотря на немного худшие характеристики методов (судя по опубликованным данным), основанных на использовании N-грамм, по сравнению со словарными методами, эти методы более предпочтительны ввиду их большей мобильности для малых ЭВМ.

Наряду с известными N-граммными методами автором предложен и реализован статистический метод использования N-грамм, который основан на подсчете частот N-грамм в словах текстов документов с учетом позиционности N-граммы. В этом случае задача обнаружения сводится к поиску слов текста документа с редкими или отсутствующими N-граммами. Так как одна ошибка в слове (например, замена буквы) ведет сразу к N неправильным N-граммам, то вероятно предпола-

гать, что хотя бы одна из них будет иметь малую вероятность появления либо будет отсутствовать вообще. Насколько известно автору методов, использующих такой подход, пока не существует.

На основе предложенного метода разработана диалоговая система обнаружения орфографических ошибок в текстах документов, функционирующая в рамках АСНТИ СО РАН. Система функционирует в интерактивном режиме, выдавая на экран терминала страницу текста документа и устанавливая курсор в первую позицию неопознанного слова, и предлагает пользователю одну из следующих возможностей:

— заменить (отредактировать) неправильное слово на правильное и обучиться на нем;

— оставить слово без изменений (новое слово) с обучением системы на нем;

— оставить слово без изменений (аббревиатура, фамилия автора и др.) без обучения.

Отличительной чертой этой системы является ее мобильность на персональных ЭВМ ввиду малого использования оперативной памяти (128 Кбайт), что достигается оригинальным представлением N-грамм (N=3) в памяти ЭВМ.

Экспериментальная апробация системы проводилась на отечественных базах данных ВИНИТИ. Визуальный анализ результатов экспериментов показал высокую степень обнаружения ошибок 90%.

В заключение доклада автор хотел бы подчеркнуть, что описанный статистический подход в задачах автоматизированного отбора (классификации) и обнаружения орфографических ошибок в текстах документов пригоден для практического применения, хотя теоретическое обоснование такого подхода нельзя считать полным. Несмотря на это, описанные системы функционируют в и продолжают развиваться.