Адаптивный лингвистический классификатор для ИПС.

Волкова И.А., Проскурня М.О.
Факультет Вычислительной Математики и Кибернетики МГУ
volkova@cs.msu.su, max@axofiber.org.ru

Ключевые слова: информационный поиск, классификация

Данная статья посвящена исследованию и реализации алгоритма адаптивной классификации (категоризации) текстов на естественном языке для информационно-поисковых систем (ИПС). В работе делается акцент на проблеме адаптации лингвистических ресурсов к новым словам и понятиям. Рассматриваемый метод автоматизированной классификации базируется на выделении именных групп с помощью предсинтаксического анализа и синтаксического РСП-анализатора, составлении семантического образа текста и процедуре определения степени близости к заданным семантическим группам. При обнаружении неизвестных словоформ морфологический анализатор строит возможные гипотезы, которые могут быть отфильтрованы на основе подходящей схемы сетевой грамматики. Все новые именные группы, состоящие из уже известных словоформ, заносятся в базу данных в виде гипотез с вероятными семантическими характеристиками, унаследованными от исходного документа, и вероятными ассоциативными связями. Выделенные нераспознанные конструкции и их компоненты (с вероятными морфологическими признаками) также заносятся в БД. Далее пользователь производит корректировку БД, по возможности уточняет все неопределённые характеристики и удаляет ошибочные записи и связи.

Чисто статистические методы информационного поиска при анализе документа строят характеристический вектор, состоящий из значимых терминов документа. Причём вес термина — его значимость — определяется, например, по совокупности частот его встречаемости в самом документе и в целом в группе документов (например, индекс TFxIDF). Для флективных языков в лучшем случае применяется морфологический модуль, который позволяет произвести нормализацию словоформ и снизить информационную избыточность. При этом в качестве ограничений взаимного расположения выделенных терминов в документе могут служить только структурные признаки: термины находятся на расстоянии меньше заданного, в одном предложении или в одном абзаце. Ясно, что точность классификации методом опорных векторов (SVM, [5]) будет сильно зависеть от репрезентативности обучающего набора. То есть документах достаточной длины показано, что SVM-классификатор даст хороший результат. Но короткие тексты, содержащие новые, не входящие в обучающий набор термины, могут быть неверно обработаны.

Рассматриваемый в данной работе комплекс является развитием лингвистического программного комплекса (ЛПК), описанного в [3]. Конечной целью проекта является разработка автоматизированного способа адаптации классификатора информационно поисковой системы к изменениям в естественном языке в рамках исследования алгоритмов поиска информации. Причём ключевым моментом является сочетание статистических методов с лингвистическими.

Архитектурно ЛПК состоит из следующих компонентов: модуль предсинтаксического анализа (RPEU), синтаксический анализатор на основе обобщённых моделей управления (ОМУ-анализатор, [4]), модуль семантических отношений (SU), модуль морфологии (RMU, [3]) и модуль классификации (TCU). При разработке данной системы получила дальнейшее развитие идеология многоуровневой клиент-серверной архитектуры: пользовательский www-клиент ↔ www-сервер + серверные сценарии ↔ ЛПК. Благодаря web-технологиям значительную часть операций с ЛПК пользователь может выполнять через глобальную компьютерную сеть. Взаимодействие компонентов происходит с помощью сетевого соединения, а для представления структурированных данных используется формат XML.

Известно, что выделение из текста именных групп с помощью синтаксического анализатора повышает точность классификации и поиска. Та как естественный язык постоянно

развивается, необходимо использовать адаптивные алгоритмы, примером которых может являться ОМУ-анализатор. Но это не окончательное решение, потому что ОМУ-анализатор может адаптировать новые модели управления только для слов с известными морфологическими признаками. Поэтому при появлении в предложении неизвестного слова адаптивный анализ будет не полным. И в итоге классификация и поиск теряют свою точность. Поэтому требуется автоматизированный метод пополнения морфологической базы. Для этого была проведена модернизация морфологического компонента, была расширена его БД собственными именами и аббревиатурами, а также реализованы следующие алгоритмы: определение характеристик неизвестных слов (построение гипотез), определение опечаток (ошибок) в словах, автоматизированное пополнение. Как отмечалось в [3], в RMU реализована переработанная модель, основанная на результатах работы [1]. Для определения ошибочных слов реализован тот же алгоритм, что и в [1], а механизм пополнения и анализа новых слов был немного обновлён: исправлены таблицы диагностических префиксов и суффиксов с учётом новых парадигматических классов и подклассов. Кроме того, реализован метод потокового пополнения из файла, когда уже известны все грамматические переменные, кроме имени П-класса.

Алгоритм работы рассматриваемого адаптивного классификатора развивает идею SVM-методов. Предлагается включать в характеристический вектор документа не только отдельные термины, но и терминологические словосочетания (именные группы, содержащие весомые термины). Важным компонентом ЛПК является модуль семантических отношений, который функционально представляет собой «псевдотезаурус». Главное отличие от настоящего тезауруса заключается в том, что SU хранит не только занесённую экспертом «достоверную» информацию, но и автоматически сформированные гипотезы. Также следует уточнить, что адаптация системы происходит при условии заранее частично заполненной базы SU. Например, в данной реализации в качестве «первого приближения» выступает упрощённый тезаурус (только с ассоциативными и иерархическими отношениями) по нескольким проблемным областям.

Построение гипотез при анализе новых текстов производится с использованием комбинированного структурно-лингвистического метода. Для наиболее весомых с точки зрения структуры документа терминологических словосочетаний, находящихся в титуле или заголовках документа, строятся и заносятся в базу данных SU вероятные ассоциативные связи. Кроме того, исследуется окружение конструкций так называемых глаголов-связок: «является», «называется», «состоит из», «основан на», «известный как», «представляет собой» и т.п. Одни и те же группы могут порождать набор различных гипотез, вследствие недетерминированного алгоритма синтаксического разбора ОМУ-анализатора, ядром которого является РСП-анализатор [2]. Кроме того, из-за вхождения в текст неизвестных слов морфологический анализатор может ещё больше расширить спектр предполагаемых ассоциаций.

Все сформированные гипотезы хранятся в базе SU до очередного сеанса контроля БД экспертом. При входе в систему ему выдаётся список новых гипотез с указанием частоты, с которой они были «подтверждены» (выделены повторно) при анализе новых текстов. Если гипотеза с точки зрения эксперта верна, он при необходимости доопределяет её семантические признаки одновременно с морфологическими признаками её составляющих (заносит в БД морфологического модуля новые слова). Ошибочные гипотезы помечаются таковыми для предотвращения построения повторного ошибочного предположения. И если такие связи впоследствии более не «подтверждаются», они автоматически удаляются из базы SU.

Ссылки:

[1] Волкова И.А. Адаптация и обучение системы общения с ЭВМ на естественном языке. Автореф. дисс. к.ф.-м.н. — М., Изд-во МГУ, 1982.

[2] Волкова И.А., Головин И.Г. Синтаксический анализ фраз естественного языка на основе сетевой грамматики. ДИАЛОГ'98, Труды межд. семинара. — М., 1998, с. 438-447.
[3] Волкова И.А., Проскурня М.О. Программный комплекс для лингвистической обработки текстов на русском языке. ДИАЛОГ'2002, Труды межд. семинара. — М., 2002, с. 96-99.
[4] Одинцев Н.В. Синтаксический анализатор на основе обобщенных моделей управления. ДИАЛОГ'02, Труды межд. семинара. — М., 2002.
[5] Boser B., Guyon M., Vapnik V. A Training Algorithm for Optimal Margin Classifier. COLT, pp 144-152

Irina A. Volkova, Maxim O. Proskurnya

Adaptive linguistic classifier for IR-systems

The article describes the research and implementation of the adaptive text-categorization algorithm for information retrieval systems. The paper makes accent on the new phrases and concepts adaptation problem. The given method of automated text-categorization is based on noun-phrases extraction with the help of pre-syntactic and EFST-syntactic analyzer, semantic image construction and semantic group correlation measurement function. After new text is processed all extracted phrases with their probabilistic semantic and morphological characteristics are stored in database. Then expert should revise the database to define all unrecognized properties, delete erroneous records and associations.