Технология смыслового анализа и поиска информации КЛЮЧИ К ТЕКСТУ

М.Г. Крейнес, А.А. Афонин, А.В. Тихонов

Рассмотрены особенности технологии смыслового анализа и поиска информации КЛЮЧИ К ТЕКСТУ, характер ее использования в экспериментах ROMIP и проблемы организации подобных экспериментов.

1. Введение (технология КЛЮЧИ К ТЕКСТУ - К^{2}Т)

Существующие объемы текстовой (в том числе, гипертекстовой) информации в электронной форме делают абсолютно нереальным личное знакомство человека с каждым текстом. Это определяет исключительную актуальность разработки информационных технологий, практически не требующих участия специалиста на этапе поиска необходимой информации и ее смысловой классификации. Традиционные методы информационного поиска по ключевым словам часто не приводят к отбору интересных текстов и отсеву неинтересных. Причина этого кроется не только в сложности для человека формирования в виде небольшого по объему списка слов адекватного его желаниям поискового образа. Недостаточно эффективно само использование в качестве критерия отбора информации просто наличия в ней определенных слов, включенных в поисковый образ. Проблематичен и альтернативный подход - априорная не ориентированная на конкретного пользователя смысловая индексация текстов, среди которых выполняется поиск.

На основании оригинальной двухуровневой модели понимания и интерпретации текстовой информации (знаковый - семиотический уровень и семантический уровень, требующий непременного участия человека для содержательной интерпретации текста) нами предложено использовать для преодоления рассмотренных трудностей принципиально новые алгоритмы вычислительного

синтеза семиотического образа текста. Основной особенностью этих алгоритмов является то, что они не используют информацию о смысле и значении слов, в частности нет нужды в смысловых тезаурусах. Предлагаемые алгоритмы в ходе формальной процедуры выделяют своеобразное «семиотическое поле» - множество слов, наиболее сильно связанных в конкретном анализируемом тексте в смысле предложенного М.Крейнесом статистического критерия, на основании сопоставления анализируемого текста с представительной для предметной области совокупностью текстов. Оказалось, что именно этот набор слов в их совокупности воспринимается человеком в качестве носителя основной содержательной и смысловой нагрузки конкретного документа.

Базой для нашей технологии являются алгоритмы построения для любого текста «смыслового» портрета – множества слов, семиотически наиболее сильно связанных между собой в конкретном анализируемом тексте. Слово «смыслового» заключено в кавычки не случайно. Хотя при анализе и интерпретации человеком получаемых в результате вычислений списков слов очевидна их осмысленность и связь с тематикой, содержанием и смыслом исследуемого текста, никакой семантической информации и знаний о грамматике языка в ходе вычислений не требуется. Для выявления семиотически связанных слов применяется предложенная М.Г. Крейнесом оригинальная метрика, использующая в качестве исходной информации только данные о комбинаторной статистике словоупотребления в анализируемом тексте и в некоторой совокупности текстов, представительной для языка, на котором написан анализируемый текст. Выбор такой референтной совокупности текстов равносилен формулировке позиций, с которых человек хочет воспринимать конкретный текст. Такой выбор можно ограничить текстами определенной группы носителей языка, например, профессиональной или политической. Задание референтной совокупности можно трактовать как неявное задание варианта (подмножества) языка, адекватного воспринимающему текст субъекту.

Таким образом, построение смыслового портрета основано на двух принципиальных гипотезах:

1. Семиотические характеристики (семиотические связи слов в тексте) являются определяющими для семантики текста.
2. Для понимания смысла конкретного текста необходимо определить совокупность текстов, в контексте которых следует воспринимать конкретный текст.

По существу, это практически фольклорные аксиомы в среде лингвистов, филологов и психологов. Достаточно вспомнить две классических формулировки:

- человек – это стиль,
- человек – это текст.

Справедливость сформулированных гипотез подтверждается весьма высокой эффективностью вычислительного анализа текстов в технологии КЛЮЧИ ОТ ТЕКСТА.

Наша технология предполагает также, что следует идентифицировать различные словоформы каждого слова (например, одного существительного в различных падежах). Такая идентификация дает возможность абстрагироваться от конкретных грамматических форм слов при построении смысловых портретов текста. Для этого используются знания о языке, на котором написан текст. В реализованном нами варианте технологии такое распознавание (так называемая лемматизация) основано на специфическом морфологическом анализе, который позволяет с достаточно большой надежностью распознавать различные словоформы конкретных слов данного языка. На сегодня такой морфологический анализ реализован для русско- и англоязычных текстов.

Рассмотренные процедуры построения смыслового портрета текста решают задачу адаптивного к интересам конкретного носителя языка (профессиональной или политической группы, индивидуума, определенного автора, издания, группы изданий) вычислительного смыслового индексирования текстовой информации.

Результаты такого вычислительного индексирования интересны сами по себе, как средство автоматического создания вторичных информационных ресурсов - списков ключевых слов, адекватно с точки зрения конкретного читателя отображающих содержание и смысл текста. Одновременно, смысловые портреты позволяют выделять в тексте наиболее важные для тематики и содержания всего текста фрагменты, что обеспечивает автоматическую генерацию рефератов. Наконец, появляется возможность вычислительной смысловой классификации текстов. Для этого предложена и используется специальная мера смысловой близости, основанная на вышеупомянутой метрике семиотической связанности слов в тексте. Предложены также способ оценки степени принадлежности конкретного текста к тематической категории, описанной набором текстов, способ классификации коллекции текстов на тематически (содержательно) однородные

группы без априорно определенного перечня таких групп и критерии и способы определения схожести корпусов текстов между собой.

Сегодня технология K^2T представляет из себя вычислительную систему с собственной структурой данных и параллельными механизмами поиска информации. K^2T реализуется на высокопроизводительных вычислительных системах.

2. Использование K^2T в экспериментах дорожки классификации

При решении задач классификации K^2T может применяться в двух принципиально различных режимах:

- априори заданы классы текстов в виде обучающих коллекций,
- классы текстов априори не определены и следует не только классифицировать коллекцию, но и сформировать представления о существующих в рамках коллекции содержательно различных классах документов.

Ввиду крайне низкого качества исходных материалов экспериментальной коллекции в экспериментах РОМИП нам пришлось сочетать оба подхода. Поскольку обучающие выборки для конкретных тем не прошли содержательного контроля, была реализована следующая процедура улучшения качества обучающих выборок. Для всей обучающей выборки была организована процедура вычислительной классификации текстов при априори не определенном числе и составе классов. В результате такой вычислительной классификации были выделены определенные тематически связанные группы текстов. Эти группы были по составу сопоставлены с обучающими выборками для конкретных тем. Для каждой из построенных выборок, отыскивали максимально схожую с ней по составу обучающую выборку для конкретной темы, затем отыскивалось пересечение построенной нами выборки с найденной максимально схожей обучающей выборкой. Эти пересечения использовались в дальнейшем в качестве «улучшенных» обучающих выборок. В Приложении приведены списки документов, вошедших в «улучшенные» обучающие выборки для оцененных в ходе эксперимента этого года тематических категорий.

Процедура классификации тестовой выборки основывалась на вычислении меры принадлежности каждого текста к обучающим коллекциям для «улучшенных» обучающих выборок, в качестве

результата классификации выбирали 3 категории, для которых значения меры принадлежности были наибольшими (таких документов оказалось около 8 тыс.). Для документов, которые не удалось классифицировать указанным способом, использовали классификацию по неулучшенным обучающим выборкам (еще 4 тыс. документов).

Для выполнения вычислений была построена база данных, содержащая информацию о частотных словарях и семиотических портретах всех проанализированных документов.

3. Проблемы организации эксперимента и предложения по его совершенствованию

Первостепенными проблемами являлись низкое качество исходной информации (текстовая, нетекстовая, в непонятной кодировке) и неоправданные надежды на возможность формирования обучающих выборок без тщательного содержательного анализа документов. Появившиеся в последний момент описания некоторых рубрик для оценщиков документов абсолютно не соответствовали реальному содержанию обучающих выборок.

По-видимому, именно эти проблемы и должны решаться в первую очередь при продолжении экспериментов. Подготовку материалов для экспериментов следует проводить серьезно и, возможно, с использованием специальных программных средств.

Для разумного анализа уже полученных результатов представляется целесообразным вычисление показателей качества классификации для обучающих выборок и их сравнение с показателями качества классификации.

К сожалению, режим анонимности не позволил организаторам предоставить участникам наиболее интересную для дальнейшей работы по улучшению технологий информацию о поведении конкретных систем – участников для каждой из конкретных категорий. Ведь только сравнивая сильные и слабые стороны технологий, которые проявляются при работе с определенными темами, характеризующимися различной структурой исходных данных, можно осознанно двигаться вперед.

Приложение

В данном Приложении приведены списки «улучшенных» обучающих выборок (для пояснений см. текст выше). После номера тематической категории следует список отнесенных к ней документов.

106
ander.htm
art-produce.htm
baraban4.htm
dis-vlad.htm
kompass1.htm
napoval.htm
tcfreestyle.htm

113
abed.htm
affa3.htm
afghan.htm
alastochkin.htm
amgonta.htm
amstaff.htm
anzhelina.htm
ashbi.htm
best-friends.htm
bullmastiff.htm
djef.htm
golden-sobaka.htm
goldybox.htm
greyhaund.htm
grief.htm
krokodiloff.htm
kuzinkerry.htm
russiancats.htm
senb-ur.htm
tchunya.htm

111
aliceflowers.htm
dovgan-centre.htm
elit-galand.htm
e-torg.htm
kbnkristal.htm
lsp44.htm

159
airis.htm
arbuzka.htm
cubanarva.htm
daylight-zone.htm
marginal-group.htm
niferx.htm
opexzvukozapis.htm
superroot.htm
zavety.htm

168
adbelkin.htm
deepspace.htm
kol2000.htm
pmob.htm
thermonuclear.htm
zviozdi.htm

177
healthy-family.htm
imigracija.htm
investor-work.htm
probig.htm
ritajobs.htm
vit-bmk1.htm

182
econom-geograf.htm
el-13-99.htm
ivt-vlgu.htm
maxxigruppa.htm
mehmath.htm
studaktiv.htm

192
avtodor.htm
nabockoff.htm
nbaslam.htm
nskvolley.htm
sportlinks.htm
vuzvolley.htm
zgia-volley.htm

194
bars-pushkino.htm
doma2003.htm
karate1936.htm
kkvb.htm
koronka.htm
ladomir1.htm
samu-rai.htm
seyken.htm

197
aerobica-class.htm
aerofitness.htm
cssc-rosto.htm
dance-kuntsevo.htm
e-darts.htm