ОЦЕНКА РЕЗУЛЬТАТОВ ПОИСКА СЕМАНТИЧЕСКИ БЛИЗКИХ СЛОВ В ВИКИПЕДИИ *

А. А. КРИЖАНОВСКИЙ*

Санкт-Петербургский институт информатики и автоматизации РАН

СПИИРАН

14-я линия ВО, д. 39, Санкт-Петербург, 199178

<aka@iias.spb.su>

УДК 681.3

Крижановский А. А. **Оценка результатов поиска семантически близких слов в Википедии** // Труды СПИИРАН. Вып. 5. — СПб.: Наука, 2007.

Аннотация. *Классификация метрик и алгоритмов поиска семантически близких слов в тезаурусах WordNet, Роже и энциклопедии Википедия расширена адаптированным HITS алгоритмом. С помощью экспериментов в Википедии оценены метрика Резника, адаптированная к Википедии, и адаптированный алгоритм HITS. Предложен ресурс для оценки семантической близости русских слов.* — Библ. 5 назв.

UDC 681.3

Krizhanovsky A. A. **Evaluation experiments on related terms search in Wikipedia** // SPIIRAS Proceedings. Issue 5. — SPb.: Nauka, 2007.

Abstract. *The classification of metrics and algorithms search for related terms via WordNet, Roget's Thesaurus, and Wikipedia was extended to include adapted HITS algorithm. Evaluation experiments on Information Content and adapted HITS algorithm are described. The test collection of Russian word pairs with human-assigned similarity judgments is proposed.* — Bibl. 5 items.

1. Введение

Под *семантически близкими словами (СБС)* подразумеваются слова близкие по значению, встречающиеся в одном контексте. Это могут быть синонимы *(чертог, дворец),* антонимы *(запутать, распутать)* и др. Во многих задачах умение составить список СБС, либо сравнить слова и вычислить - какие слова ближе по значению, оказывается востребованным.

Во-первых, это так называемый «поиск по смыслу», при котором пользователь вводит слово *мобильник,* но видит страницы, содержащие другие слова, например, *мобильный телефон, сотовый* и др. Поисковая система расширила или переформулировала запрос с помощью СБС.

Во-вторых, запросно-ответные системы на этапе обработки вопроса пытаются вычислить, к какой области относится вопрос пользователя, пытаются найти похожие вопросы в базе данных. Поиск вопросов основан, в том числе, и на использовании списков СБС.

В-третьих, для выбора одного из значений многозначного слова, (например, слово *граф* может обозначать либо титул, либо математический объект) используют СБС.

* Полная версия технического отчета: <http://arxiv.org/abs/0710.0169>.

• Исследования, результаты которых представлены в настоящей работе, были частично выполнены в рамках проектов РФФИ # 05-01-00151 и # 06-07-89242, проекта # 16.2.35 программы «Математическое моделирование и интеллектуальные системы» Президиума РАН, проекта # О-1.9 программы «Фундаментальные основы информационных технологий и компьютерных систем» ОИТВС РАН.

В-четвёртых, есть интерес к автоматическому созданию специальных словарей - тезаурусов на основе СБС. Прелесть таких тезаурусов в том, что они строятся по тексту и могут наглядно, в виде картинки, предъявить ключевые понятия, найденные в тексте, и то, как они связаны.

В-пятых, трудоёмкая задача составления словарей синонимов (и не только синонимов) требует кропотливой работы лексикографов. Своевременную помощь оказывают поисковые алгоритмы, предлагающие списки близких по значению слов для последующего вдумчивого разбора лингвистом.

Количество научных работ, посвящённых Википедии, стремительно растёт.[1] Осветим одну из граней этого феномена, а именно: корпус текстов Википедии обладает особой привлекательностью для поисковых алгоритмов. Вики занимает нишу между, с одной стороны, размеченными корпусами текстов, а с другой - интернет-страницами (где нет никаких надёжных подсказок для алгоритмов, кроме гиперссылок и частоты слов). Перечислим «изюминки» викитекстов с точки зрения машинной обработки:

- заголовок, максимально точно соответствующий теме текста;
- первый абзац обычно даёт краткое описание термина, может содержать основные ключевые слова;
- наличие внутренних ссылок на статьи по данной теме; специальный раздел ссылок *«Смотри также»;*
- специальный формат для ссылок на статью о том же термине на другом языке (интервики);
- категории, классифицирующие документы по их тематической принадлежности.

Системы поиска семантически близких слов в Википедии помогут пользователям, во-первых, находить энциклопедические статьи, близкие по тематике к заданным, что позволит более глубоко изучить исследуемое понятие. Во-вторых, помогут в указании недостающих ссылок между связанными по смыслу статьями.

2. Эксперименты: метрика Резника и категории Википедии[2]

Далее описаны результаты и особенности вычисления метрики Резника, адаптированной к Википедии, и результаты работы алгоритма AHITS. Алгоритм AHITS (адаптированный HITS) описан в работе [2].

Резник в работе [4] предложил считать, что два слова тем более похожи, чем более информативен концепт *(Information Content),* к которому соотнесены оба слова, то есть чем ниже в иерархии находится общий верхний концепт. Например, для категорий Википедии *Лётчики* и *Самолёты* ближайшим общим концептом будет *Авиация* (рис. 1). Информативность концепта было предложено считать на основе частотности термина в корпусе текстов: чем более частотен терм (и его подклассы), тем меньшей информативностью он обладает [4].

В работе [5] метрика Резника была адаптирована к Википедии и информативность категории вычислялась как функция от числа гипонимов категорий[3], а не статистически:

[1] См. <http://en.wikipedia.org/wiki/Wikipedia:Wikipedia_in_academic_studies>.

[2] Эксперименты можно повторить с помощью программы Synarcher версии 0.12.4, см. *Release Notes* в программе, доступной по адресу: <http://synarcher.sourceforge.net>.

Труды СПИИРАН. 2007. Вып. 5. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 5. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

$$res\ hypo(c_1, c_2) = \frac{\log(hypo(lcs(c_1, c_2)) + 1)}{\log(C)}$$

где *lcs* — ближайший общий родитель концептов c_1 и c_2 (от англ. *least common subsumer), hypo* — число гипонимов этого родителя, а *C* — общее число концептов в иерархии. На рис. 1 видно, как уменьшается число подкатегорий и статей при спуске по иерархии вниз *(hypo* — первое число в скобках). Информативность категории $res\text{h}_{\text{ypo}}$ при этом увеличивается (второе число). Таким образом, можно вычислить семантическую близость слов. Например, для *дирижабля* и *самолёта* семантическое сходство равно 0.575, поскольку ближайшим общим концептом будет концепт «Воздушные суда».

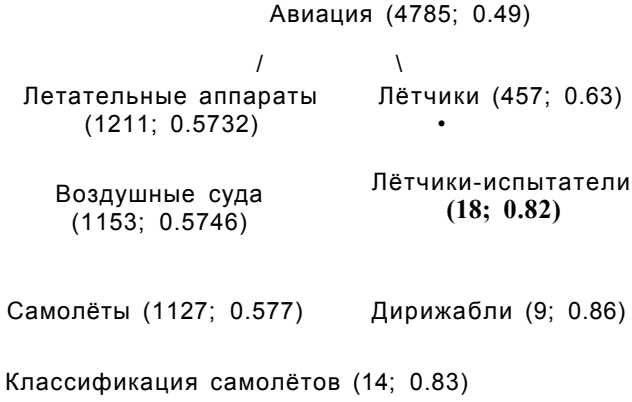

Рис. 1. Фрагмент иерархии категорий в Википедии.

Эксперименты по вычислению метрики resh_{ypo} в википедиях на английском, simple[4] и русском языках показали, что есть некоторые особенности, определяемые структурой Википедии: в графе категорий есть циклы и стоит задача выбора корневой категории. Метрика resh_{ypo} рассчитана на дерево без циклов, но это не так в Википедии, что было учтено при реализации данной метрики. Для оценки метрик и алгоритмов, вычисляющих близость значений слов, использовался тестовый набор из 353 пар английских слов (далее 353-TC).[5]

Корреляция *AHITS* и resh_{ypo} с тестовыми данными были рассчитаны с помощью программы Synarcher. Экспериментальные данные о результатах поиска СБС в тезаурусах WordNet, Роже и английской энциклопедии Википедия для других метрик и алгоритмов взяты из работы [5], данные по алгоритму ESA[6] из [3]. Классификация метрик и алгоритмов поиска СБС, предложенная в [5], расширена адаптированным *HITS* алгоритмом. Таким образом, можно выделить подходы, основанные на учёте:

[3] Гипонимы категории *K* в Википедии - это все подкатегории К, а также все статьи, принадлежащие этим подкатегориям и категории К.

[4] Простая Английская Википедия, см. <http://simple.wikipedia.org>.

[5] Данные доступны по адресу: <http: //www.cs.technion.ac.il/~gabr/resources/data/wordsim353/wordsim353.html>.

[6] В *ESA* концепт - это название энциклопедической статьи Википедии. На вход подаются два текста. По ним строятся два вектора из концептов Википедии. Для сравнения текстов сравнивают два вектора, например, с помощью косинусного коэффициента.

Труды СПИИРАН. 2007. Вып. 5. ISSN 2078-9181 (печ.), ISSN 2078-9599 (онлайн)
SPIIRAS Proceedings. 2007. Issue 5. ISSN 2078-9181 (print), ISSN 2078-9599 (online)
www.proceedings.spiiras.nw.ru

1. *расстояния в таксономии,* лучшее значение корреляции с тестовой коллекцией 353-TC равно 0.48 у метрики *lch* для английской Википедии (ВП);
2. *анализа веб-ссылок* — 0.38-0.39 у *AHITS* алгоритма для английской ВП;
3. *частотности слов в корпусе* — 0.72 алгоритм *ESA* [3] для английской ВП;
4. *совпадения (перекрытия) текстов* — 0.21, метрика *lesk* для WordNet.

Оценка корреляции результатов поиска СБС с тестовым набором 353-TC показала, что алгоритм AHITS даёт несколько лучший результат (0.38-0.39), чем адаптированная метрика Резника (0.33-0.36) на данных английской Википедии.

3. Заключение и тестовый набор русских слов

Для оценки поиска семантически близких слов проведены эксперименты, которые показали, что наилучшую точность даёт алгоритм *ESA* [3], учитывающий частотность слов в корпусе текстов. Разрабатываемый *AHITS* алгоритм выполняет поиск на основе анализа гиперссылок и категорий. Предлагается расширить *AHITS* алгоритм учётом частотности слов. Это позволит обрабатывать документы без ссылок (или с малым их числом).

Капица П. Л. писал: «...теория — это хорошая вещь, но правильный эксперимент остаётся навсегда» [1]. Однако, чтобы провести эксперимент и оценить результаты поиска близких по значению слов, нужен тестовый набор (эталон), который создан людьми вручную, а не автоматически.

Для английского языка такой набор есть - это 353 пары слов, в оценке которых участвовало два десятка людей. Именно этот набор использовался и для оценки корреляция *AHITS* алгоритма и метрики res_{hypo} с тестовыми данными. Итого уже более десяти метрик и алгоритмов были оценены с помощью этого набора, но только для данных на английском языке.

Чтобы выполнить оценку работы алгоритмов в русской Википедии, необходим тестовый набор из русских слов. Такой набор создан, следующий этап — это простановка вручную семантической близости парам слов добровольцами.[7]

Литература

1. *Капица П. Л.* Эксперимент, теория, практика. М., 1974. 288 с.
2. *Крижановский А. А.* Автоматизированное построение списков семантически близких слов на основе рейтинга текстов в корпусе с гиперссылками и категориями // Компьютерная лингвистика и интеллектуальные технологии. Труды международной конференции «Диалог 2006». Бекасово, 2006. С. 297-302.
3. *Gabrilovich E., Markovitch S.* Computing semantic relatedness using Wikipedia-based Explicit Semantic Analysis // In Proceedings of The 20th International Joint Conference on Artificial Intelligence (IJCAI). Hyderabad, India, January, 2007. // <http: // www.cs.technion.ac.il/~gabr/papers/ijcai-2007-sim.pdf> (по состоянию на 3.11.2007).
4. *Resnik P.* Disambiguating noun groupings with respect to WordNet senses // In Proceedings of the 3rd Workshop on Very Large Corpora. MIT, June, 1995. // <http: // xxx.lanl.gov/abs/cmp-lg/9511006> (по состоянию на 3.11.2007).
5. *Strube M., Ponzetto S.* WikiRelate! Computing semantic relatedness using Wikipedia // In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI 06). Boston, Mass., July 16-20, 2006. // <http: // www.eml-research.de/english/research/nlp/publications.php> (по состоянию на 3.11.2007).

[7] См. страницу проекта <http://ru.wikipedia.org/wiki/Участник:AKA_MBG/Wordsim>.