Меры семантической близости статей Википедии и их применение к обработке текстов

П.Е. Велихов

Аннотация. Рассмотрены меры семантической близости статей Википедии и их применение к задачам обработки текстов и информационного поиска. Приведены требования к мерам семантической близости для задач вычисления близости между парой статей и для ранжирования всех статей Википедии относительно заданной. Предложены эвристические методы эффективного ранжирования для отдельного класса мер. Приведены экспериментальные данные, подтверждающие эффективность предложенного подхода. Кратко рассмотрены методы, реализованные в системе Texture.

Ключевые слова: семантическая близость, Википедия, анализ текстов, информационный поиск.

Введение

В последнее время в сети интернет стремительно развиваются открытые базы данных, такие как Wikipedia [1], Wiktionary [2], dmoz.org [3] и другие. Семантическая информация, накопленная в этих базах, позволяет существенно улучшить методы автоматической обработки текстов и информационного поиска. Вместо ключевых слов, обычно используемых в информационно-поисковых системах, становится возможным использование конкретных терминов предметной области. Википедия является ведущим источником среди выше перечисленных и в настоящее время английская Википедия включает в себя: более 2,5 миллиона статей, среди которых порядка 300,000 статей-категорий и 700,000 многозначных терминов; более 1 миллиона статей перенаправления; и более 60 миллионов ссылок с одной статьи Википедии на другую. Такой объем данных в совокупности с широким покрытием предметных областей и достаточно глубоким материалом энциклопедического стиля, рецензированием сообществом редакторов Википедии, предоставляет возможность использовать Википедию как семантический словарь для автоматической обработки текстов. К тому же Википедия постоянно обновляется, существующие статьи пополняются последними событиями, появляются новые разделы и тексты. Но, в отличие от структурированных словарей глаголов, таких как VerbNet [4] и FrameNet [5], которые применяются для анализа текста, Википедия – это в первую очередь словарь фразеологизмов, не покрывающий глаголы. В свою очередь статьи Википедии не содержат семантические структуры, присущие словарям глаголов, поэтому требуются другие методы работы с семантикой. Одним из таких методов является мера семантической близости.

В данной статье мы рассмотрим современные меры семантической близости и их применения в контексте системы Texterra [6], разработанной в Институте системного программирования РАН. На примере системы Texterra становится очевидным, что мера семантической близости должна быть эффективной с вычислительной точки зрения, что дисквалифицирует целые семейства популярных в литературе функций.

1. Система анализа текстов Texterra

Целью разработки системы Texterra является использование Википедии для улучшенного поиска и навигации в текстовых базах данных. Пользуясь семантикой Википедии, Texterra, в отличие от традиционных информационно-поисковых систем, оперирует более богатым представлением текстовых документов, что в свою очередь позволило нам реализовать новые возможности для этого класса систем. Модель документа в Texterra представляется как множество фразеологизмов из текста и, так как фразеологизмы берутся из Википедии, между фразами определено семантическое расстояние. При этом многозначные фразеологизмы разрешаются в однозначные на основе меры близости [7]. При наличии качественной функции семантического подобия, такая модель позволяет определять доминантные фразы документа [8], производить рубрикацию текстов без использования обучающих выборок, помогать реформулировать запрос пользователя и вычислять фасеты при навигации по результатам полнотекстового запроса и по рубрикатору.

1.1. Архитектура системы

На Рис. 1 приведена архитектура системы. Сначала, в стадии предварительной обработки из Википедии извлекается: словарь фраз, включающий в себя статьи перенаправления; словарь многозначных фраз с конкретными вариантами значений и граф ссылок Википедии.

1.2. Анализ текста

Процесс анализа текста состоит из следующих этапов:

1. Лексический анализ.
2. Выделение частей речи.
3. Морфологический анализ.
4. Выделение фразеологизмов Википедии.
5. Разрешение многозначности.
6. Выделение доминантных фраз.

В качестве лексического анализатора системы используется пакет OpenNLP [9], который также используется для определения частей речи. Морфологический анализ английского языка состоит из самых простых правил, нормализующих существительные единственного и множественного числа. Далее следует стадия выделения фраз Википедии. На этой стадии выделяется максимально длинная последовательность лексем, являющаяся названием статьи Википедии, с несколькими ограничениями: термин должен представлять собой следующую стандартную последовательность частей речи [32]:

((Adj|Noun)+|(Adj|Noun)(NounPrep)?(Adj|Noun)*)Noun*

После этой стадии разрешаются многозначные фразеологизмы. Как упоминалось ранее, Википедия включает в себя словарь многозначных фраз, причем общее количество таких фраз составляет более чем 700000. Это обусловлено тем, что в Википедии представлены многие редкие значения сокращений и варианты, похожих по смыслу терминов. Например, термин «Platform» имеет 17 различных значений, среди которых: «Railway platform», «Computing platform», «Political platform» и «Oil platform». Для разрешения такого рода многозначности используется алгоритм на основе меры семантической близости. Для каждого многозначного фразеологизма система собирает несколько однозначных, обнаруженных на предыдущей фазе фраз, в дальнейшем именуемыми как контекст многозначного фразеологизма, и вычисляет семантическую близость между контекстом и каждым значением многозначной фразы. Выбирается значение, максимизирующее семантическую близость к контексту. Алгоритм разрешения многозначности детально рассмотрен в секции 7. Последняя стадия анализа текста – это выделение доминантных терминов текста. Так как текст на этой стадии представляется множеством фраз в пространстве меры семантической близости, система использует центральности фраз и выбирает фразы с высокой центральностью. Метод определения центральности фраз также рассмотрен в секции 7.

1.3. Рубрикация, поиск и навигация по обработанным текстам

После анализа текста, система Texterra позволяет проводить рубрикацию текстов относительно линейного или иерархического рубрикатора, где каждая рубрика задается названием статьи Википедии. Из-за широкого охвата Википедии, почти любая интересующая пользователя рубрика будет представлена какой-нибудь

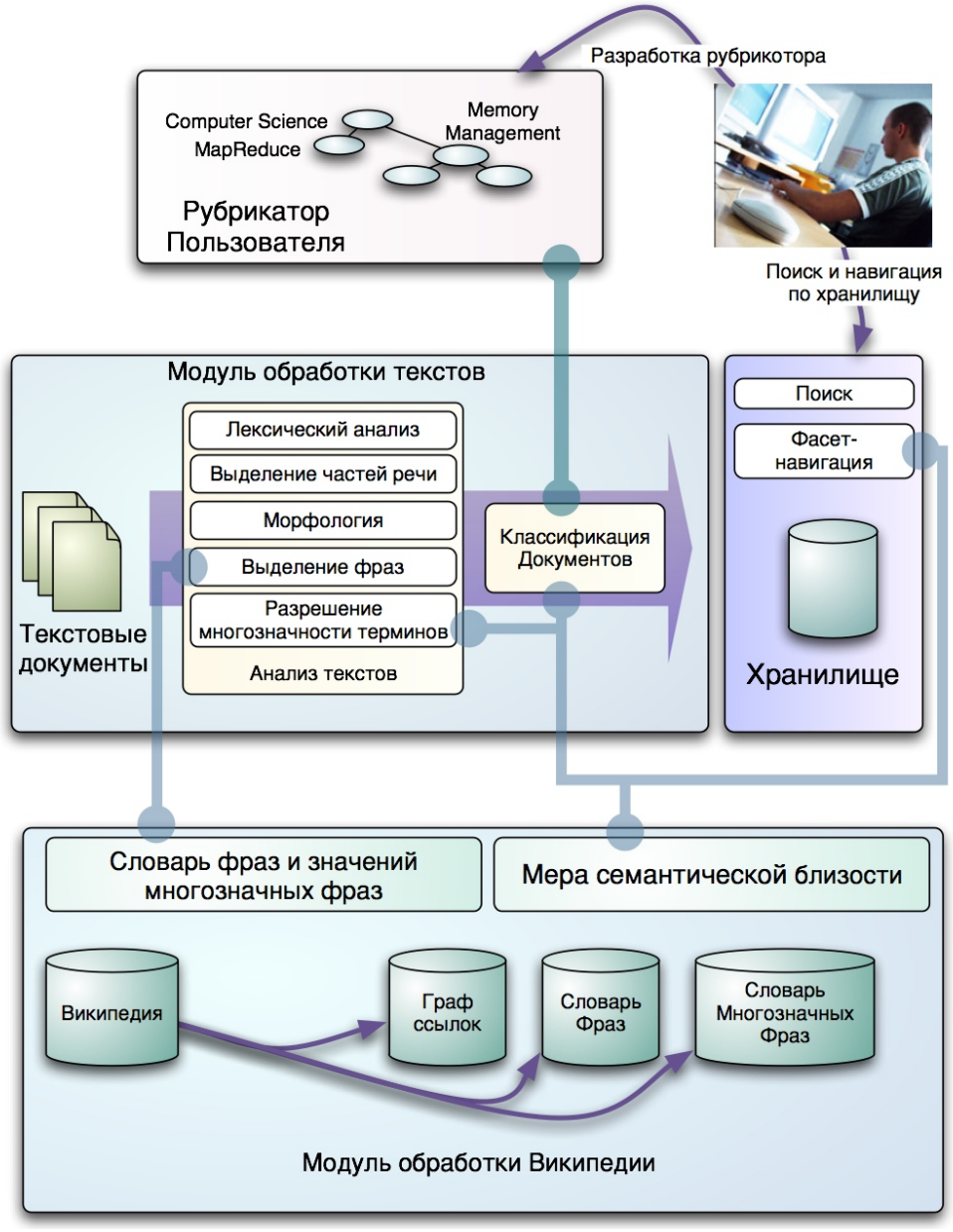

Рис.1. Архитектура системы Texterra

статьей. Рубрикация производится автоматически на основе меры семантической близости и рассмотрена более детально в разделе 7. Поиск по документам выполняется в стиле полнотекстового запроса, широко применяемого в текстовых базах данных, но система автоматически предлагает расширить запрос похожими терминами или сузить запрос с помощью техники фасет-навигации. Обе эти техники основаны на ранжировании статей Википедии по мере семантической близости к фразам запроса и подробно описаны в разделе 5.

1.4. Вычислительные требования к функции семантической близости

Очевидно, мера семантической близости применима в большом количестве задач автоматической обработки текстов и в системах поиска и навигации. Но именно из-за такого широкого спектра применений встает вопрос о

вычислительной сложности функции близости. Например, алгоритм разрешения многозначности фраз обращается к функции $|c|*|m|$ раз, где $|c|$ – количество фраз в контексте, а $|m|$ - количество разных значений фразы. В среднем в Википедии многозначный термин имеет 17 значений, размер контекста в алгоритме разрешения многозначности составляет 7 фраз и около 40% терминов в произвольном публицистическом тексте являются многозначными. Таким образом, небольшой текст в 1000 слов может потребовать порядка 3000 вычислений функции близости. Рубрикация текстов требует подсчета семантической близости от каждого термина текста до каждой рубрики в рубрикаторе и сравнима по сложности с задачей разрешения многозначности. Далее, при выделении доминантных терминов текста используются методы центральности, которые требуют вычисления семантической близости между всеми парами фразеологизмов. А при расширении запросов и вычислении фасетов, используется ранжирование всех статей Википедии относительно терминов запроса или конкретной рубрики. Таким образом, вычислительные требования к функции семантической близости очень высоки в контексте практических систем обработки текстов.

2. Меры семантической близости

В последние несколько лет было предложено большое количество мер семантической близости для статей Википедии и Веб-страниц. Но основной упор делался на качество результатов и эстетике различных мер, а подробного анализа вычислительной сложности мер проведено не было. Далее мы рассмотрим несколько классов предложенных мер и проведем анализ сложности на двух задачах: вычисление семантической близости двух статей и ранжирование статей Википедии относительно заданной. При этом в задаче ранжирования нам необходимы статьи, наиболее близкие к заданной. Например, при расширении запроса пользователя, система должна предложить ему небольшое количество близких вариантов и не обязательно производить ранжирование всех статей Википедии. Вместо этого, можно ограничиться несколькими десятками результатов, что дает нам дополнительные возможности для оптимизации вычисления ранжирования.

2.1. Контентные и ссылочные меры

Традиционно в информационно-поисковых системах близость текстовых документов вычисляется как угол между векторами документов, образуемыми весами ключевых слов документов по схеме TF-IDF, изначально введенной Джонсом в [10]. Меру семантической близости статей Википедии, основанную на TF-IDF авторы предложили в [11] и [12]. Но из-за того, что такого рода меры используют только ключевые слова без фразеологизмов и игнорируют информацию, содержащуюся в ссылках, эти меры страдают от недостатка точности, как показано Менцером в [13]. Вычислительная сложность этой меры для задачи вычисления близости двух статей пропорциональна количеству ключевых слов этих статей и достаточно эффективна для практического использования. Но задача ранжирования статей по данной мере, даже при условии, что выбираются первые несколько десятков семантически похожих документов, не может быть эффективно решена, так как развернутая статья Википедии будет содержать сотни ключевых слов, и будет семантически похожа почти на любую другую статью с ненулевым весом. В задаче информационного поиска, где ранжируются документы по мере близости к запросу, похожая проблема решается методом порогового алгоритма, например алгоритма Фагина [14], где система может вычислить первые десятки результатов запроса эффективно, используя только первые записи упорядоченного инвертированного списка. Но при наличии сотен ключевых слов в запросе, такой метод перестает быть эффективным.

Таким образом, меры, основанные на контенте, не обладают достаточной вычислительной эффективностью и качеством для их использования в практических системах.

С другой стороны, ссылочная структура Википедии очень богата и, в основном, содержит ссылки высокого качества, которые высоко релевантны для меры семантической близости. В среднем, каждая статья Википедии содержит около 25 ссылок, причем большие информативные статьи содержат на порядок больше: до 1200 ссылок в статье «United States». В основ-

ном, ссылки указывают на близкие по смыслу статьи, а маргинальные неинформативные ссылки исключаются модераторами Википедии. Кроме того, в Википедии есть специальные ссылки, несущие информацию о семантической близости: в конце статей Википедии часто присутствует секция «См. Также» в которой содержатся ссылки на похожие по смыслу статьи; кроме того статьи содержат ссылки на категории Википедии – тем самым, обозначая примерную область статьи в таксономии Википедии. В связи с такой богатой и информативной структурой ссылок Википедии, а также из-за негативных свойств мер, основанных на контенте статей, мы используем меру близости, основанную только на ссылках Википедии.

2.2. Обзор ссылочных мер семантической близости

Меры близости, основанные на ссылках Википедии, легче всего разделить на три класса: меры парного случайного блуждания, меры случайного блуждания и нерекурсивные меры, такие как косинус, меры Дайса и Джакарта и прочие. Последние меры часто использовались в области информационного поиска.

2.2.1. Меры парного случайного блуждания

В это семейство попадают популярные рекурсивные функции семантической близости: SimRank [15] и мера близости вершин Ньюмана [16]. Обе меры основаны на предположении, что похожие вершины графа ссылок ссылаются на вершины, которые в свою очередь похожи. Оба алгоритма рекурсивны, и в начальном шаге рекурсии все вершины похожи только на себя с коэффициентом 1. Далее в рекурсии вершины графа, связанные с похожими вершинами становятся более похожими друг на друга со значением, зависящим от степеней вершин и коэффициента затухания рекурсии. Ниже приведена формула SimRank для итеративного подсчета:

$$S_{ij} = \frac{C}{k_i k_j} \sum_{u,v} A_{iu} A_{vj} S_{uv}$$

В формуле S_{ij} обозначает ячейку матрицы подобия вершин, A_{ij} ячейку матрицы смежности, k_i и k_j – степени вершин i и j, а C – коэффициент затухания. Эту формулу можно также интерпретировать, как математическое ожидание пути с учетом затухания, которое пройдут два синхронизованных блуждающих объекта перед тем, как встретиться в одном и том же узле, если они стартовали в вершинах i и j. Эта модель основана на алгоритме PageRank [19], где релевантность узла определяется вероятностью нахождения в произвольный момент в заданной вершине графа.

Ньюман в [16] предложил использовать похожую меру близости, но с более прямой вероятностной интерпретацией и отсутствием некоторых артефактов, присущих SimRank. Предложенная им мера подобия измеряет отношение количества путей в графе между двумя вершинами к ожидаемому количеству путей двух случайных вершин: чем больше путей между двумя вершинами, тем более вершины похожи. Но с другой стороны, любые две вершины большой степени связаны большим количеством путей, поэтому мера Ньюмана использует отношение наблюдаемого количества путей к ожидаемому количеству путей в случайном графе. Эта мера задается следующей формулой:

$$S_{ij} = \delta_{ij} + \frac{2m}{k_i k_j} \sum_{l=1}^{\infty} \alpha^l \lambda_1^{-l+1} \left[A^l\right]ij$$

В ней δ_{ij} – дельта-символ Кронекера, $2m/k_ik_j$ (m – количество ребер графа) – обратное значение ожидаемого количества путей длинной 1, l – длина пути, a – параметр затухания, λ_l – максимальный собственный вектор матрицы смежности A.

Здесь ожидаемое количество путей длины l приблизительно вычисляется формулой $k_ik_j/2m*\lambda_l^{l-1}$. Мера Ньюмана приводит к более качественным результатам, чем мера SimRank, так как она свободна от нескольких дефектов последней: SimRank учитывает только пути четной длины и не компенсирует результат ожидаемым количеством путей в графе, что приводит к тому, что длинные пути компенсируют результат и приходится полностью полагаться на коэффициент затухания C, чтобы ослабить влияние длинных путей на конечный результат.

Обе меры этого класса весьма устойчивы к шуму, т. е. к ссылкам плохого качества и имеют

убедительную интерпретацию в модели парного случайного блуждания. Но, к сожалению, вычислительная сложность этих мер очень высока. Хотя количество итераций, за которое достигаются качественные результаты небольшое, например авторы SimRank рекомендуют 5-6 итераций, из-за очень маленького диаметра Википедии обе меры практически не вычислимы. Диаметр основного связанного компонента Википедии составляет 4,5 ребер, но из самых значимых статей все вершины доступны менее чем через 4 ребра. Таким образом, оба алгоритма вычисляют полную матрицу семантической близости, а вычислительная сложность алгоритмов, даже при оптимизации алгоритмов, как в [17], составляет $O(n^3)$, где n – количество ребер Википедии. Форагас в своей работе [18] предложил использовать метод Монте-Карло для приблизительного вычисления меры SimRank, но полученные им результаты на порядки отличались от точной меры, и такая техника пригодна только для приблизительного ранжирования.

2.2.2. Меры случайного блуждания

Модель случайного блуждания успешно используется для ранжирования Веб-страниц в поисковых системах интернета, например, широко известен алгоритм PageRank [19], основанный на этой модели. Оливьер в своей работе [20] провел сравнительный анализ мер семантической близости статей Википедии, основанных на модели случайного блуждания. Он рассмотрел несколько мер, основанных на мере Грина, локальный PageRank, индекс перекрестной цитируемости и косинус с точки зрения качества ранжирования этих мер. Мера Грина определяется следующей формулой:

$$G_{ij} = \sum_{t=0}^{\infty} \delta_{ij} P^t - \nu_j$$

В этой формуле P – это стохастическая матрица построенная на основе матрицы смежности, где $P_{ij} = 1/\sum A_{ik}$, ν – собственный стационарный вектор матрицы P. Данная мера соответствует времени, проведенному блуждающим пользователем в узле j, если он стартовал из узла i. Похожая на меру Грина, мера локального PageRank выражается следующей формулой:

$$S_{ij} = \left(\sum_{t=0}^{\infty} c(1-c)^t P^t \right)_{ij}$$

В этой формуле случайный процесс блуждания стартует из узла i, с равной вероятностью переходит по ссылкам графа и с вероятностью c возвращается в узел i. Без коэффициента c, S_i превращается в стационарный вектор цепи Маркова ν.

По наблюдениям Оливьера, качественные результаты дают только модификации меры Грина, а локальный PageRank из-за очень малого диаметра Википедии превращается в глобальный PageRank, не несущий информации о семантической близости. Самой удачной мерой в нашей работе оказалась мера $S_{ij} = G_{ij} log(\nu_j)$; на данных, полученных ручным ранжированием статей Википедии, она дала наилучшие результаты.

Похожий подход использовал Ли, в разработанной им мере близости PageSim [21].

По вычислительной эффективности меры случайного блуждания существенно выигрывают у мер парного блуждания, но при малом диаметре Википедии использование этих мер означает обход всего графа ссылок Википедии, как для задачи подсчета близости двух статей, так и для ранжирования. То есть вычислительная сложность данного семейства мер для обеих задач составляет $O(n)$, где n – количество ссылок Википедии. Такая эффективность недостаточна для применения этих метрик в практических разработках, подобных системе Texterra.

2.2.3. Нерекурсивные меры

Последнее семейство мер – это простые, нерекурсивные меры близости, которые традиционно применялись в области информационного поиска. Эти меры – косинус, мера Дайса и Джакарта, определенные на векторах смежности Википедии. Мера косинуса имеет следующий вид:

$$S_{ij} = \frac{A_i \bullet A_j}{\|A_i\| \|A_j\|},$$

а две другие меры удобнее выражаются в множественно-теоретической форме:

$$\text{мера Дайса: } \frac{|N_i \cap N_j|}{|N_i| + |N_j|},$$

мера Джакарта: $\frac{\left|N_i \cap N_j\right|}{\left|N_i \cup N_j\right| - \left|N_j \cap N_j\right|}$,

в обеих формулах N_i обозначает множество вершин смежных с вершиной i. Самой привычной мерой в области информационного поиска является косинус, но при использовании в качестве меры близости вершин графа она теряет свой смысл меры угла в векторном пространстве и поэтому две другие меры более привлекательны. Проведение детальной оценки качества мер требует ручного ранжирования статей Википедии, поэтому мы положились на оценку качества, полученную Оливьером, где мера косинуса показала достаточно хорошие результаты, близкие к мере Грина. Две другие нерекурсивные меры дают похожие результаты, по субъективным оценкам нами была выбрана мера Дайса. Она имеет простую интерпретацию для вершин графа, как отношение количества общих соседей вершин к сумме степеней вершин. По этой причине мы выбрали меру Дайса для системы Texterra. Вариацией меры Дайса также пользуется Мильне в разработанной им информационно-поисковой системе [22].

Проведем анализ вычислительной сложности мер этого семейства. Все меры нерекурсивного семейства имеют похожую вычислительную сложность. Задача вычисления семантической похожести двух статей имеет сложность $O(d)$, где d – наибольшая степень статьи, при допущении, что ссылки Википедии предварительно отсортированы и пересечения и объединения считаются в один проход. Хотя в Википедии встречаются статьи со степенью (числом входящих и выходящих ссылок) более 120 тысяч, такая сложность все равно на порядок эффективнее, чем перебор всего графа ссылок Википедии.

Иначе обстоит дело с задачей ранжирования. Чтобы получить точное ранжирование статей Википедии по одной из вышеперечисленных метрик от исходной вершины i, надо рассмотреть все вершины, отстоящие на два ребра от i, и отсортировать этот список по значению меры. Таким образом, задача ранжирования имеет неприемлемую сложность $O(n^3 log(n))$. Но, как упоминалось ранее, для практических задач нам требуются первые результаты ранжирования, а не весь список. И если допустить некоторую неточность в результатах ранжирования, то возможно решить эту задачу на несколько порядков эффективнее, пользуясь статистическими особенностями графа ссылок Википедии. В следующем разделе мы рассмотрим статистические свойства ссылок Википедии, а затем представим несколько эвристических алгоритмов ранжирования по нерекурсивным мерам.

3. Статистические свойства графа ссылок Википедии

Граф ссылок Википедии является частным случаем так называемых безмасштабных или комплексных графов, которые детально изучались, начиная с конца 90-ых годов. У таких графов есть несколько важных статистических свойств в сравнении с обычными случайными графами Эрдоша-Реньи. Во-первых, в безмасштабных графах степени вершин распределены по степенному закону, в отличие от экспоненциального распределения в случайных графах. Вероятность, что случайный узел графа будет иметь степень k равна: $P(k)=ck^{-\gamma}$. Из такого распределения следует большое количество узлов с очень большой степенью, так называемый длинных хвост распределения. Это распределение порождает безмасштабное свойство графов: вне зависимости от степени конкретного узла, в любом окружении с одной и той же вероятностью обнаружатся узлы с большими и меньшими степенями вершин. Вершины с большой степенью в безмасштабных графах называются хабами, и выполняют связывающую функцию. Именно из-за этих вершин диаметр безмасштабных графов очень мал, и составляет $log(log(n))$, где n - количество узлов графа [23].

3.1. Модели образования безмасштабных графов

Самая известная модель образования безмасштабных графов была предложена Барабаши в [24], где Барабаши предложил принцип предпочтительного соединения, как основной механизм формирования безмасштабного графа. В отличие от модели Эрдоша-Реньи, где узлы соединяются друг с другом с равной вероятностью, в модели Барабаши новые узлы присоединяются к существующим с вероятностью пропорциональной степени последних

узлов. На примере Википедии это означает, что, скорее всего в новой статье появятся ссылки на известные статьи с большим количеством входящих ссылок, чем на малоизвестные. Более формально, в модели Барабаши в каждую единицу времени t в граф добавляется узел t с m новыми ребрами, смежными с существующими вершинами, где вероятность присоединения к вершине i равна:

$$\Pr(i,t) = \frac{k_i}{mt}$$

Модель Барабаши генерирует безмасштабные графы со степенным распределением степеней вершин, однако она упускает несколько важных свойств Википедии и графа ссылок Веб-страниц: во-первых, в Википедии статьи с большим количеством входящих ссылок также имеют большое количество исходящих ссылок, и распределение исходящих ссылок также происходит по степенному закону; во-вторых, страницы Википедии ссылаются не на случайные страницы, а на страницы с похожим содержанием, что приводит к более кластеризированной структуре графа. Здесь под кластеризацией мы понимаем условную вероятность, что две вершины i и j имеют общее ребро, если существует вершина k, имеющее общие ребра с i и j. Эти два свойства Википедии, а особенно высокая кластеризация, оказываются принципиально важными для разработки эффективных алгоритмов ранжирования, которые будут рассмотрены в следующем разделе.

Менцер в своей работе [25] также заметил, что ссылки в Википедии и в Веб-страницах ссылаются на похожие страницы, и предложил смешанную модель образования графов ссылок. По аналогии с моделью Барабаши, вероятность соединения узла t с существующим узлом i равна:

$$\Pr(i,t) = \alpha\frac{k_i}{mt} + (1-\alpha)\left(\frac{1}{\sigma_c(i,t)} - 1\right)^{-\varphi}$$

В этой модели первый член – это предпочтительное соединение по модели Барабаши, а второй член соединяет новые вершины с похожими, где $\sigma_c(i,t)$ – похожесть статей или веб страниц по косинусу угла TF-IDF векторов их текстового содержания. Модель Менцера гораздо точнее описывает граф Википедии, включая коэффициент кластеризации, чем модель Барабаши, и дает достаточно близкое представление процесса формирования Википедии и ее статистических свойств.

3.2. Измерения параметров Википедии

Далее мы приводим основные статистические параметры Википедии, замеренные нами, которые согласуются с исследованиями Википедии проведенными в [26] и [27]. Первый параметр – это параметр распределения степеней вершин γ в Википедии примерно равен 2,2, что согласуется с измерениями этого параметра в графах Веб-страниц и других безмасштабных графах. Случайные выборки статей Википедии приведены на Рис. 2 и Рис. 3. На втором рисунке отображены степени входящих и исходящих ссылок, а на третьем – степени исходящих ссылок.

Последний параметр Википедии, имеющий особо важное значение, – это коэффициент кластеризации, и его среднее значение в Википедии составляет $7*10^{-3}$, что примерно на порядок превышает ожидаемый коэффициент кластеризации в безмасштабном графе, построенном по модели Барабаши. Но этот коэффициент согласуется с моделью Менцера, что в свою очередь обозначает, что похожие статьи Википедии находятся в

Рис. 2. Распределение степеней статей Википедии (входящие и исходящие ссылки)

Рис. 3 Распределение количества исходящих ссылок

самом ближайшем окружении исходной статьи. Следовательно, нерекурсивные функции семантической близости показывают результаты сравнимого качества с более сложными рекурсивными мерами.

4. Приблизительное ранжирование статей Википедии по нерекурсивным мерам близости

В предыдущем разделе мы рассмотрели статистические свойства графа ссылок Википедии, где одним из основных является высокий коэффициент кластеризации. Используя это свойство Википедии, мы разработали эвристические алгоритмы, которые на порядки улучшают вычислительную сложность ранжирования, внося небольшие погрешности в результат ранжирования.

Как описывалось ранее, при ранжировании результатов обычно требуются первые, самые близкие по мере близости статьи. Рассмотрим задачу вычисления первых k результатов для нерекурсивной меры семантической близости Дайса. При ранжировании статей относительно заданной статьи a, первые k результатов будут иметь большое количество ссылок на те же статьи, на которые существует прямая ссылка с a. Это особенно верно для статей высокой степени. В этом случае, из-за высокого коэффициента кластеризации графа ссылок Википедии, вероятность того, что между a и первыми k результатами будет прямая ссылка очень высока. Рассмотрим конкретный пример ранжирования относительно статьи «United Kingdom»: статья «Labour Party» оказывается 20-ой в списке наиболее семантически близких статей по мере Дайса. При этом у этих двух статей около 3000 общих соседей в графе ссылок. Вероятность того, что существует прямая ссылка с «United Kingdom» на «Labour Party» составляет $1 - (1 - c)^{3000}$, где c – коэффициент кластеризации Википедии, равный $7*10^{-3}$. Таким образом, вероятность наличия прямой ссылки при таком количестве общих соседей равна 1. Это наблюдение позволило нам разработать эвристические методы ранжирования, описанные далее.

4.1. Алгоритм ВС

Первый эвристический метод называется ВС (все ссылки), в котором мы ранжируем только те статьи, которые связаны с исходной статьей прямой ссылкой. Этот алгоритм дает результат высокого качества относительно точного ранжирования, однако для статей с высокой степенью ссылок он недостаточно эффективен. Вычислительная сложность алгоритма ВС составляет $O(d^2)$, где d – максимальная степень исходной статьи и всех с ней связанных статей. Для статей, как «United States», алгоритму ВС придется отранжировать более 100000 статей, вычисляя при этом меру близости для каждой из них. Таким образом, в худшем случае алгоритм ВС менее эффективен, чем ранжирование по мерам случайного блуждания.

4. 2. Алгоритм ИС

Проблема алгоритма ВС – это статьи с большим общим количеством ссылок, входящих и исходящих. Но в статьях Википедии на порядок меньше исходящих ссылок, чем входящих, что очевидно на Рис. 2 и Рис. 3. Для статей с достаточно большим количеством исходящих ссылок, можно ранжировать только те статьи, на которые ссылается исходная статья. Этот метод уступает в качестве результатов эвристике ВС, но его вычислительная сложность составляет $O(od*d)$, где od – количество исходящих ссылок исходной статьи. Так как статьи с большой степенью имеют ограниченное количество исходящих ссылок, этот метод на порядок эффективнее, чем ВС.

4.3. Алгоритм ИС + первые k

Эвристический метод, который используется в системе Texterra – это улучшенный алгоритм ИС. Он основан на гипотезе, что статьи, похожие друг на друга, будут в свою очередь ссылаться на похожие статьи. При ранжировании статей относительно статьи q, мы сначала ранжируем все статьи, на которые имеется исходящая ссылка с q. Затем мы добавляем все исходящие ссылки с первых k отранжированных статей и заново ранжируем расширенный список статей. Вычислительная сложность этого метода также составляет $O(od*d)$, но при этом при небольшой константе k, например $k=20$, этот алгоритм добивается более качественных результатов, чем алгоритм ВС.

Сравнительная таблица вычислительной сложности различных мер семантической бли-

зости приведена ниже, а в следующем разделе приведены результаты экспериментов для оценки точности и эффективности эвристических методов ранжирования.

Вычислительная сложность различных мер семантической близости

Класс мер	Близость 2-х статей	Ранжирование
Парное блуждание	$O(n^3)$	$O(n^3)$
Случайное блуждание	O(n)	O(n)
Нерекурсивные, точное ранжирование	O(d)	$O(n^3)$
Нерекурсивные, ВС	O(d)	$O(d^2)$
Нерекурс. ИС+ k	O(d)	O(od*d)

5. Экспериментальные результаты

Мы провели серию экспериментов, чтобы оценить качество и эффективность эвристических методов ранжирования. В первом эксперименте мы оцениваем качество ранжирования относительно алгоритма точного ранжирования. Случайная выборка из Википедии плохо подходит для данного эксперимента, так как, в основном, будут выбраны статьи с малым количеством ссылок. Поэтому мы разбили статьи Википедии по степеням на интервалы в 200 степеней. Максимальная степень статей в нашем эксперименте составляет 20000 из-за низкой эффективности алгоритма точного ранжирования.

При сравнении результатов эвристических алгоритмов с точным методом, мы оценивали процент первых 20-ти результатов ранжирования по эвристикам, которые также оказались в первых 20-ти результатах точного ранжирования. Результаты эксперимента приведены на Рис. 4.

В следующем эксперименте мы вычислили количество статей, которые приходится ранжировать каждым методом. В этом эксперименте мы считали дубликаты статей, и в некоторых случаях количество статей немного завышено. Результаты этого эксперимента приведены на Рис. 5.

Эти эксперименты показывают, что эвристические методы ранжирования дают достаточно точные результаты, при этом на два порядка эффективнее справляются с задачей ранжирования. Таким образом, мы смогли найти эффективное решение обоих задач вычисления семантической близости: как вычисления близости двух статей, так и ранжирования.

В последних двух экспериментах мы замерили время вычисления обоих задач. Для задачи вычисления меры близости между двумя статьями, мы выбрали статью «Beer», чья степень составляет 2000, и считали семантическую близость со статьями различной степени. Для второй задачи мы вычисляли первые 20 результатов ранжирования по эвристике ИС + первые 40 для статей различных степеней. Эксперимент был поставлен на компьютере MacPro с 8-ми ядерным процессором с частотой 2 гигагерц и 16-ю гигабайтами оперативной памяти. Результаты приведены на Рис.6 и Рис.7.

В заключение стоит отметить, что нерекурсивные меры семантической близости с эвристическим ранжированием – это единственные

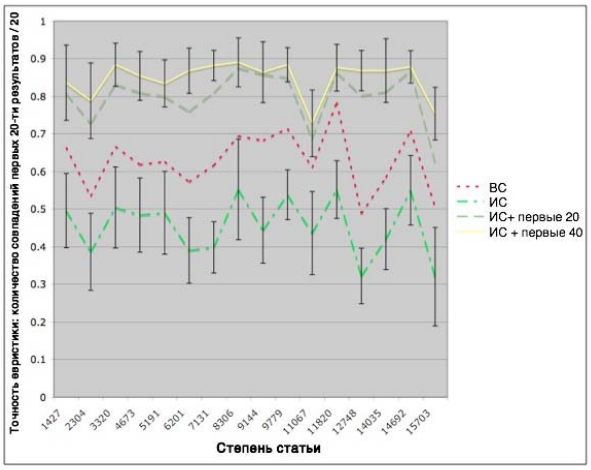

Рис.4. Точность результатов эвристического ранжирования

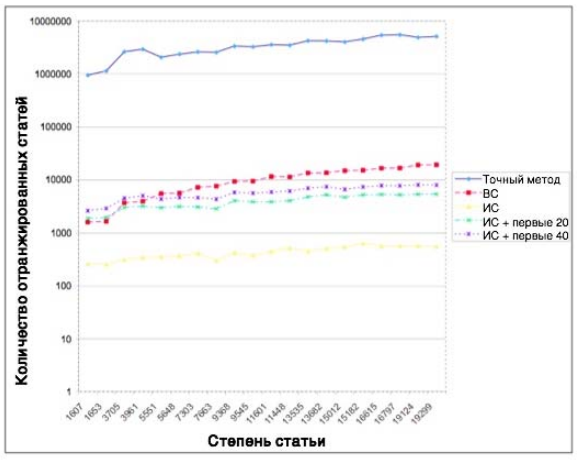

Рис.5. Вычислительная эффективность эвристического ранжирования

Рис.6. Время вычисления меры близости для пар статей

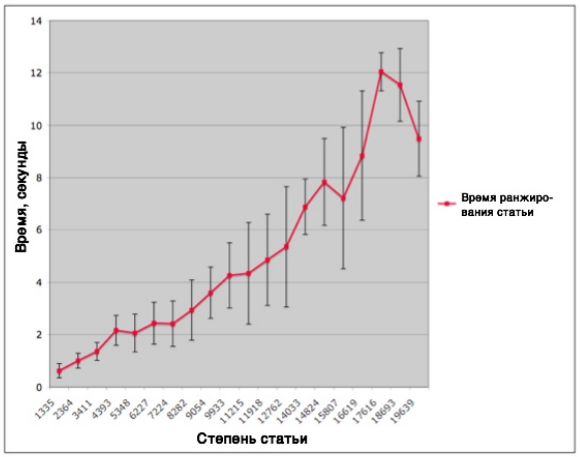

Рис.7. Вычислительная эффективность ранжирования

приемлемые для практического применения меры из предложенных выше. При этом качество результатов этих мер достаточно высоко и мало уступает более сложным мерам. Если принять во внимание модель образования Википедии, где ссылки с большой вероятностью указывают на похожие по смыслу статьи, что в свою очередь приводит к высокой кластеризации статей, то этот результат становится понятным: близкие по смыслу статьи должны находиться в ближайшем окружении исходной статьи, а сложные алгоритмы блуждания с большой вероятностью замкнутся на этом высоко-кластеризированном окружении.

6. Применение меры семантической близости для задач обработки текста и информационного поиска

Далее рассмотрим применения меры семантической близости для следующих задач: обработки текстов, а точнее разрешение многозначности фразеологизмов Википедии, встреченных в тексте; рубрикации текстов; расширение запросов при информационном поиске и реализации фасетной навигации по рубрикатору и результатам запросов.

6.1. Разрешение многозначных фразеологизмов

Как упоминалось ранее, Википедия содержит большое количество многозначных терминов. К настоящему моменту общее количество терминов в статьях английской Википедии составляет примерно 700000 терминов, а среднее количество значений многозначного термина равно 17. При этом практически все сокращения и очень многие часто упоминаемые термины оказываются многозначными. Например, такие часто используемые слова как «Platform», «Function», «History», «Machine» являются многозначными, так как есть различные толкования этих терминов в разных предметных областях. Такого рода многозначность терминов приводит к тому, что в произвольном тексте около 40% терминов оказываются многозначными. При этом необходимо выбрать самое подходящее значение многозначного термина, так как от этого очень сильно зависит качество всех последующих методов обработки текста и алгоритмы поиска, классификации и навигации текстов.

В системе Texterra мы используем простой алгоритм разрешения многозначности, основанный на мере семантической близости. Алгоритм работает следующим образом: в окружении многозначного термина t выделяется контекст из нескольких однозначных терминов $c_1 \ldots c_n$; мы выбрали контекст длиной в 7 терминов. Далее, из значений термина $t_1 \ldots t_k$ выбирается значение, максимизирующее сумму близости значения и терминов контекста:

$$t_{result} = \arg\max_{t_i} \sum_{j=1}^{n} sim(t_i, c_j)$$

Для оценки точности этого алгоритма мы использовали примеры разрешения многозначности из самой Википедии. В статьях Википедии ссылки имеют вид: *[подпись|ссылка]*, где *подпись* – это текст, который виден читателю статьи, а *ссылка* – это полное название статьи Википедии. Мы выбрали из Википедии ссылки, у которых подпись совпадает с многозначным термином Википедии, а ссылка указывает на одно из значений. Далее мы немного модифицировали меру близости для этой задачи: для каждой статьи мы использовали вес, обратный степени статьи. Затем проверили алгоритм на описанном множестве, где он правильно определил значения многозначных терминов в 72% случаев. При этом в 90% случаев правильное значение попадает в первые три самых близких к контексту значений. Результат в 72% дает достаточно хорошие результаты для задач классификации текстов, информационного поиска и фасет-навигации. Неправильно определившиеся термины привносят лишь небольшие неточности в другие алгоритмы, так же основанные на мере семантической близости. Для более сложных задач, как например извлечение фактов из текстов, требуется разработка более точных методов разрешения многозначности.

6.2. Выведение доминантных фраз текста

После выделения фразеологизмов Википедии в тексте и разрешения многозначных терминов модель документа представляет собой множество терминов Википедии в пространстве меры семантической близости. В этом пространстве есть группы сильно связанных терминов, которые близки по смыслу друг к другу; встречаются и термины, удаленные от всех других терминов документа. Обычно, маргинальные термины представляют собой несущественные термины документа или даже ошибки алгоритмов выделения терминов и разрешения многозначности. Но при наличии меры семантической близости, мы можем посчитать так называемую центральность терминов текста и присвоить соответствующие веса терминам документа. Эти веса могут быть использованы в алгоритмах рубрикации текстов, информационного поиска, вычисления фасетов и так далее.

В теории графов используются несколько мер центральности, самые известные из них – это центральность по кратчайшим путям и центральность по собственному значению. Центральность по кратчайшим путям оценивает процент кратчайших путей, проходящих через каждый узел графа, относительно общего количества кратчайших путей:

$$C_B(v) = \frac{1}{n^2} \sum_{s,t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

В приведенной формуле переменная σ_{st} обозначает количество кратчайших путей из вершины s в вершину t, а $\sigma_{st}(v)$ – количество кратчайших путей проходящих через вершину v, обе переменные принимают значение 0 или 1. Мера центральности по кратчайшим путям считается самой качественной мерой и широко применяется для кластеризации графов и разбиении графа на тесно связанные подмножества [31]. Но эта мера обладает высокой вычислительной сложностью в $O(n^3)$, поэтому она применима только на графах малого объема.

Другая, часто используемая мера центральности – это центральность по собственному значению, похожая на известную меру PageRank. Граф документов преобразуется в стохастическую матрицу, в которой итерационным методом находится стационарный вектор. Так как метод итерации сходится достаточно быстро, сложность вычисления данной меры центральности составляет $O(n^2)$.

Мера центральности по кратчайшим путям дает более качественную оценку важности термина статьи и только из-за ее высокой вычислительной сложности на текстах с большим количеством терминов приходится использовать центральность по собственным значениям.

6.3. Рубрикация текстов

Автоматическая рубрикация текстов – сложная задача, которая обычно решается методами машинного обучения с использованием обучающей выборки заранее отклассифицированных текстов, как например, в системе Athena [28]. Из-за отсутствия семантики текстов, алгоритмы машинного обучения, такие как, например, метод опорных векторов, обучаются семантике из примеров классификации.

Такие системы довольно трудозатратны не только на фазе их создания, но и на фазе поддержки, так как при изменении рубрикатора или большого обновления базы документов, требуется пересмотреть и изменить обучающую выборку.

Разработав качественную меру семантической близости, можно создать автоматический рубрикатор текстов на ее основе, без применения обучающей выборки. В системе Texterra как раз реализован такой рубрикатор, в котором рубрики являются статьями Википедии. Пользователь выбирает список рубрик и в любой момент может дополнить или изменить этот список. Алгоритм классификации текстов работает по такому же принципу, как и алгоритм разрешения многозначности: текстовый документ, состоящий из выделенных терминов $t_1 \ldots t_n$, при заданных рубриках $c_1 \ldots c_k$ классифицируется в рубрику, которая максимизирует сумму близости терминов к рубрике:

$$c = \arg\max_{c_i} \sum_{j=1}^{n} sim(c_i, t_j)$$

Данный алгоритм так же без существенных модификаций поддерживает классификацию в несколько рубрик, а также иерархическую классификацию. Чтобы повысить вычислительную эффективность рубрикации, можно использовать основные термины статьи, используя меру центральности терминов, описанную выше.

Чтобы оценить качество классификации, мы провели эксперимент с новостным сайтом Google News [29], на котором новости классифицируются в несколько широких рубрик, таких как Политика, Бизнес, Наука, Спорт, и т.д. Мы выбрали соответствующие статьи Википедии в качестве рубрик для нашего алгоритма и при рубрикации получили 80% совпадения с рубрикацией Google News. При этом ошибки рубрикации оказались спорными и одинаково хорошо подходили нескольким рубрикам. Например, темы Медицина и Бизнес в Google News очень часто пересекались, так как статьи описывали разработку медикаментов фармацевтическими компаниями.

6.4. Информационный поиск

Очевидное применение меры семантической близости для информационного поиска – это ранжирование результатов запросов по этой мере, где дополнительная семантическая мера позволила бы серьезно улучшить качество результатов. К сожалению, построить эффективную систему поиска с использованием семантической близости невозможно из-за отсутствия эффективных методов индексирования документов для выполнения запросов. Как упоминалось ранее, стандартные системы полнотекстового поиска используют инвертированный список и пороговый алгоритм для существенного сокращения числа потенциально релевантных документов. При использовании меры семантической близости вся коллекция документов потенциально релевантна запросу и должна быть проранжирована. Из-за этого мера семантической близости неприменима для поиска и ранжирования результатов по релевантности. Но существуют вспомогательные применения меры близости – это реформулирование запроса и фасет-навигация по результатам запроса.

6.4.1. Фасет-навигация

Когда пользователь вводит запрос к отдельной коллекции документов, он часто получает слишком много результатов, среди которых сложно найти нужные документы. Фасет-навигация позволяет пользователю получить обзор тем полученных результатов и сузить свой запрос до нужной ему подтемы. В поисковых системах часто используют алгоритмы кластеризации для группировки результатов поиска в фасеты, как например в [30], но при этом встает задача вывода правильного названия кластера, которая пока не решается качественно. С использованием меры семантической близости, фасеты для навигации по результатам можно выбирать из терминов, наиболее похожих на запрос, но которые также встречаются в документах.

Мы реализовали простой алгоритм фасет-навигации, который работает следующим образом: при запросе пользователя q и наборе документов $d_1 \ldots d_n$, найденных полнотекстовой поисковой системой, фасеты вычисляются как пересечение терминов наиболее похожих на запрос q, с терминами из $d_1 \ldots d_n$. Из списка полученных фасетов, выбираются первые k фасетов, наиболее близкие к запросу. Такой простой ал-

горитм не всегда производит качественные фасеты, так как не учитывается смысловая близость терминов документа самому документу и фасеты могут отражать второстепенные термины. Улучшить такой алгоритм можно, предварительно вычисляя центральность каждого термина относительно других терминов документа и ранжируя фасеты по совокупности семантической близости к запросу и средней величины центральности термина в документах-результатах запроса. При таком подходе получаются более качественные фасеты, но возрастает нагрузка на предварительную обработку текстов, так как подсчет центральности требует вычисления расстояния между каждой парой терминов документа.

6.4.2. Реформулирование запроса

Более простая техника использования словаря Википедии и меры семантической близости – это реформулирование запроса. Если запрос пользователя не нашел всех нужных документов, система может посоветовать расширить запрос похожими на запрос терминами. Для этого можно взять первые k результатов ранжирования терминов запроса. Другое применение Википедии к реформулированию запроса – это обнаружение многозначных терминов в запросе и предоставление пользователю возможности выбрать точное значение термина.

Заключение

Мы рассмотрели основные семейства предложенных в литературе мер семантической близости для статей Википедии и различные приложения этих мер для обработки текстов и информационного поиска на основе системы анализа и поиска текстов Texterra. Мы выяснили, что в практических системах алгоритм вычисления меры семантической близости должен быть максимально эффективен с вычислительной точки зрения. В свою очередь, это заставило нас отказаться от сложных мер, основанных на модели случайного блуждания. Для нерекурсивных мер семантической близости задача вычисления близости между парой статей решается эффективно, а для задачи ранжирования мы разработали приближенные алгоритмы, которые позволяют достаточно эффективно решать задачу ранжирования с небольшой погрешностью. На основе выбранной нами меры мы реализовали алгоритмы разрешения многозначности фразеологизмов, вычисления доминантных фраз документа, классификации документов, вычисления фасетов для фасет-навигации и метод реформулирования запросов. Эти алгоритмы, построенные с использованием меры семантической близости, не нуждаются в обучении, что избавляет разработчиков систем от необходимости создавать и модифицировать обучающие выборки и другие параметры традиционных алгоритмов. Алгоритмы, основанные на мере семантической близости, показывают достаточно хорошие результаты и могут быть использованы в прикладных системах.

Литература

1. Wikipedia: www.wikipedia.org
2. Wiktionary: www.wiktionary.org
3. Dmoz: www.dmoz.org
4. Karin Kipper, Hoa Trang Dang, Martha Palmer. Class-Based Construction of a Verb Lexicon, AAAI-2000.
5. Baker, Collin F., Fillmore, Charles J., and Lowe, John B. The Berkeley FrameNet project. COLING-ACL 1998.
6. Texterra: a toolkit for text mining: www.modis.ispras.ru/texterra
7. D. Turdakov, P. Velikhov. Semantic Relatedness Metric for Wikipedia Concepts Based on Link Analysis and its Application to Word Sense Disambiguation. In proceedings of SYRCoDIS, 2008.
8. M. Grineva, M. Grinev, D. Turdakov, P. Velikhov, A. Boldakov. Harnessing Wikipedia for Smart Tags Clustering. In proceedings of KASW'08.
9. Open NLP: opennlp.sourceforge.net
10. Spärck Jones, Karen, A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation 1972.
11. Gabrilovich, E.; Markovitch, S. Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. Proceedings of the 20th International Joint Conference on Artificial Intelligence, 2007
12. Michael Strube, Simone Paolo Ponzetto, WikiRelate! Computing Semantic Relatedness Using Wikipedia, AAAI 2006.
13. Filippo Menczer, Combining Link and Content Analysis to Estimate Semantic Similarity. Proc. 13th Intl. WWW Conf. Alt. Track Papers and Posters, pp. 452-453, 2004.
14. Ronald Fagin, Amnon Lotem Y, Moni Naor Z, Optimal aggregation algorithms for middleware, Principles of Database Systems 2001.
15. Glen Jeh, Jennifer Widom. SimRank: a measure of structural-context similarity. Proceedings of the eighth ACM SIGKDD international conference on Knowledge

discovery and data mining, July 23-26, 2002, Edmonton, Alberta, Canada.
16. Leicht, E. A. and Holme, Petter and Newman, M. E. J., Vertex similarity in networks, Phys. Rev. E, 73:026120, 2006.
17. D. Lizorkin, P. Velikhov, M. Grinev, D. Turdakov. Accuracy Estimate and Optimization Techniques for SimRank Computation. In Proceedings of the 34th Int. Conf. on Very Large Data Bases (VLDB'08).
18. D. Fogaras and B. Racz. Scaling link-based similarity search. In WWW '05: Proceedings of the 14th international conference on World Wide Web, pages 641-650, New York, NY, USA, 2005. ACM.
19. L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web, 1998.
20. Yann Ollivier, Pierre Senellart, Finding Related Pages Using Green Measures: An Illustration with Wikipedia, In AAAI (2007).
21. Z. Lin, I. King, and M. R. Lyu. PageSim: A novel link-based similarity measure for the World Wide Web. In WI '06.
22. Milne, D., Witten, I.H. and Nichols, D.M. (2007). A Knowledge-Based Search Engine Powered by Wikipedia. In Proceedings of the ACM Conference on Information and Knowledge Management (CIKM'2007).
23. R. Cohen and S. Havlin. Scale-free networks are ultrasmall. Physical Review Letter, 90(5):058701, 2003.
24. Albert R. and Barabási A.-L. Statistical mechanics of complex networks. Rev. Mod. Phys. 74, 47–97 (2002).
25. Menczer Filippo. Evolution of document networks. Proceedings of the National Academy of Sciences of the United States of America, 101: 5261-5265, 2004.
26. V. Zlatic, M. Bozicevic, H. Stefancic, and M. Domazet, Wikipedias: Collaborative web-based encyclopedias as complex networks, Phys. Rev. E 74, 016115 (2006).
27. A. Capocci, V. D. P. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli, Preferential attachment in the growth of social networks: The internet encyclopedia Wikipedia, Phys. Rev. E 74, 036116 (2006).
28. Rakesh Agrawal, Roberto Bayardo, Ramakrishnan Srikant, Athena: Mining-based interactive management of text databases, EDBT 2000.
29. Google News: news.google.com
30. Wouter de Winter, Maarten de Rijke, Identifying Facets in Query-Biased Sets of Blog Posts, ICWSM 2007
31. M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks, Phys. Rev. E 69, 026113 (2004).
32. Juneston J.S. and Katz S.M., Technical Terminology: some linguistic properties and algorithms for identification in text. In Proceedings of ICCL-95.

Велихов Павел Евгеньевич. Научный сотрудник НИИСИ РАН. Окончил Калифорнийский университет, Сан-Диего в 2000 году. Имеет 9 публикаций. Круг научных интересов: базы данных и искусственный интеллект.