Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов

Мария Гринева, Максим Гринев
{upa|maxim}@grinev.net

Аннотация. В статье предлагается новый метод извлечения ключевых терминов из текстовых документов. В качестве важной особенности метода мы отмечаем тот факт, что результатом его работы являются группы ключевых терминов; при этом термины из каждой группы семантически связаны одной из основных тем документа. Метод основан на комбинации следующих двух техник: *мера семантической близости терминов, посчитанная с использованием Википедии*; *алгоритм для обнаружения сообществ в сетях*. Одним из преимуществ нашего метода является отсутствие необходимости в предварительном обучении, поскольку метод работает с базой знаний Википедии. Экспериментальная оценка метода показала, что он извлекает ключевые термины с высокой точностью и полнотой.

1. Введение

Ключевыми терминами (*ключевыми словами* или *ключевыми фразами*) являются важные термины в документе, которые могут дать высокоуровневое описание содержания документа для читателя. Извлечение ключевых терминов является базисным этапом для многих задач обработки естественного языка, таких как *классификация документов*, *кластеризация документов*, *суммаризация текста* и *вывод общей темы документа* [7]. В этой статье мы предлагаем метод для извлечения ключевых терминов документа, используя Википедию в качестве ресурса, насыщенного информацией о семантической близости терминов.

Википедия (www.wikipedia.org) – свободно распространяемая энциклопедия, на сегодняшний день являющаяся самой большой энциклопедией в мире. Она содержит миллионы статей, доступных на нескольких языках. В сентябре 2008 года Википедия содержит более 2.5 миллионов статей (более 6 миллионов, если считать перенаправляющие страницы, представляющие синонимы заголовка основной статьи). Обладая огромной сетью ссылок между статьями, большим числом категорий, перенаправляющих страниц (redirect pages) и страниц для многозначных терминов (disambiguation pages), Википедия представляет собой исключительно мощный ресурс для нашей работы и для

многих других приложений обработки естественного языка и информационного поиска.

В основе нашего метода лежит использование следующих двух техник: мера семантической близости, посчитанная по Википедии, и алгоритм анализа сетей, а именно, алгоритм Гирвана-Ньюмана для обнаружения сообществ в сетях. Ниже мы дадим краткое описание этих техник.

Установление семантической близости концепций в Википедии является естественным шагом на пути к построению инструмента, полезного для задач обработки естественного языка и информационного поиска. За последние три года появилось порядочное количество работ по вычислению семантической близости между концепциями с использованием различных подходов [13,14,4,19,21]. Работа [14] дает развернутый обзор многих существующих методов подсчета семантической близости концепций с использованием Википедии. Хотя метод, описываемый в нашей работе, не устанавливает каких-либо требований к способу определения семантической близости, эффективность работы метода зависит от качества работы выбранного метода подсчета семантической близости. Для экспериментов, описанных в этой работе, мы использовали метод подсчета семантической близости, описанный в работе Д. Турдакова и П. Велихова [21].

Зная семантическую близость терминов, мы можем построить *семантический граф* для всех терминов обрабатываемого документа. Семантический граф представляет собой взвешенный граф, в котором узлами являются термины документа, наличие ребра между парой терминов означает, что эти два термина семантически близки, весом ребра является численное значение семантической близости этих двух терминов. Мы заметили, что граф, построенный таким образом, обладает важным свойством: семантически близкие термины «сбиваются» в плотные подграфы, в так называемые *сообщества*, наиболее массивные и сильно связанные подграфы, как правило, соотносятся с главными темами документа, и термины, входящие в такие подграфы, являются ключевыми для данного документа. Новшество нашего подхода состоит в применении алгоритма обнаружения сообществ в сетях, который позволяет нам выявить тематические группы терминов, и затем выбрать из них наиболее плотные. Такие наиболее плотные группы терминов являются результатом работы метода – тематически сгруппированными ключевыми терминами.

Задача анализа структуры сетей и обнаружения сообществ в них на сегодняшний день хорошо изучена. Было предложено много алгоритмов, которые с успехом применялись для анализа социальных сетей [22], сетей цитирования научных статей [16, 3], сетей покупок товаров крупных Интернет-магазинов таких как Amazon [1], биохимических сетей [6] и многих других. В то же время авторам данной работы неизвестны примеры применения таких алгоритмов к сетям, построенным на основе Википедии. В нашем методе используется алгоритм, предложенный М. Ньюманом и

М. Гирваном [15]. Существуют работы, показывающие, что данный алгоритм является высокоэффективным при анализе как синтетических сетей, так и сетей реального мира.

2. Близкие работы

В области статистической обработки естественного языка существуют классические подходы к извлечению ключевых терминов: *tf.idf* и *анализ колокаций* (collocation analysis) [7]. Tf.idf (term frequency-inverse document frequency) является популярной метрикой при решении задач информационного поиска и анализа текста [17]. Tf.idf представляет собой статистическую меру того, насколько термин важен в документе, который является частью коллекции документов. С использованием Tf.idf важность термина пропорциональна количеству встречаемости термина в документе и обратно пропорциональна количеству встречаемости термина во всей коллекции документов. В то время как tf.idf используется для извлечения ключевых терминов, состоящих из одного слова, анализ коллокаций используется для обнаружения *фраз*.

Подход Tf.idf, дополненный анализом коллокаций, позволяет извлечь *ключевые фразы*. Оба подхода требуют наличия некоторой коллекции документов для сбора статистики; такую коллекцию документов называют *обучающим множеством*. Качества работы подходов зависит от того, насколько удачно подобрано обучающее множество.

Преимуществом данных подходов является простота реализации и удовлетворительное качество работы, когда обучающее множество хорошо подобрано. Благодаря этим преимуществам данные подходы широко распространены на практике. Мы бы хотели отметить интересный факт: существуют работы [9,11,2,8], где Википедия использовалась в качестве обучающего множества, и было показано, что Википедия может служить хорошим обучающим множеством для многих практических приложений.

Существует альтернативный класс подходов к решению задач обработки естественного языка (извлечение ключевых слов является одной из таких задач), и данная работа принадлежит к этому классу подходов. Подходы этого класса основаны на использовании знании о семантической близости терминов. Семантическая близость терминов может быть получена при помощи словаря или тезауруса (например, WordNet [12]), но нас интересуют работы, использующие семантическую близость терминов, полученную по Википедии.

Посчитать семантическую близость терминов с использованием Википедии можно двумя способами: используя гипертекстовые ссылки между статьями Википедии, которые соответствуют данным терминам [13,14,21], или измеряя косинус угла между векторами, построенными по текстам соответствующих статей Википедии [4]. Существует множество работ, где семантическая

близость терминов, полученная по Википедии, используется для решения следующих задач обработки естественного языка и информационного поиска: разрешение лексической многозначности термина [10,18,8,21], выведение общей темы документа [20], категоризация [5], разрешение кореферентности (coreference resolution) [19].

Авторам данной статьи неизвестны работы, где семантическая близость терминов использовалась бы для извлечения ключевых терминов документа, однако, работа [5] является наиболее близкой к нашей. В работе [5] решается задача категоризации текста, при этом из терминов текста строится семантический граф, аналогично тому, как мы предлагаем в данной работе. Идея применения алгоритмов анализа графов в этой работе проявляется в простой форме: выбираются наиболее центральные термины в графе при помощи алгоритма оценки центральности (betweenness centrality), далее эти термины используются для категоризации документа.

Мы выделяем следующие преимущества нашего метода:

- Наш метод не требует обучения, в отличие от описанных традиционных подходов. Благодаря тому, что Википедия является крупномасштабной и постоянно обновляемой миллионами людей энциклопедией, она остается актуальной и покрывает много специфических областей знаний. Таким образом, практически любой документ, большая часть терминов которого описана в Википедии, может быть обработан нашим методом.
- Ключевые термины сгруппированы по темам, и метод извлекает столько различных тематических групп терминов, сколько различных тем покрывается в документе. Тематически сгруппированные ключевые термины могут значительно улучшить выведение общей темы документа (используя, например, применение метода «spreading activation» по графу категорий Википедии, как описано в [20]), и категоризацию документа [5].
- Наш метод высокоэффективен с точки зрения качества извлеченных ключевых терминов. Экспериментальные оценки метода, обсуждаемые далее в этой статье, показали, что метод извлекает ключевые термины из документов с высокой точностью и полнотой.

3. Метод извлечения ключевых терминов

Метод состоит из следующих пяти шагов: 1) извлечение терминов-кандидатов; 2) разрешение лексической многозначности терминов; 3) построение семантического графа; 4) обнаружение сообществ в семантическом графе; 5) выбор подходящих сообществ.

3.1. Извлечение терминов-кандидатов

Целью этого шага является извлечение всех терминов документа и подготовка для каждого термина набора статей Википедии, который потенциально могут описывать его значение.

Мы разбираем исходный документ на лексемы, выделяя все возможные N-граммы. Для каждой N-граммы мы строим ее морфологические вариации. Далее для каждой из вариации производится поиск по всем заголовкам статей Википедии. Таким образом, для каждой N-граммы мы получаем набор статей Википедии, которые могут описывать ее значение.

Построение различных морфологических форм слов позволяет нам расширить поиск по заголовкам статьей Википедии и, таким образом, находить соответствующие статьи для большей порции терминов. Например, слова *drinks*, *drinking* и *drink* могут быть связаны с двумя статьями Википедии: *Drink* и *Drinking*.

3.2. Разрешение лексической многозначности терминов

На данном шаге для каждой N-граммы мы должны выбрать наиболее подходящую статью Википедии из набора статей, который был построен для нее на предыдущем шаге.

Многозначность слов – распространенное явление естественного языка. Например, слово «платформа» может означать железнодорожную платформу, или платформу программного обеспечения, а также платформу, как часть обуви.

Правильное значение многозначного слова может быть установлено при помощи контекста, в котором это слово упоминается. Задача разрешения лексической многозначности слова представляет собой автоматический выбор наиболее подходящего значения слова (в нашем случае – наиболее подходящей статьи Википедии) при упоминании его в некотором контексте.

Существует ряд работ по разрешению лексической неоднозначности терминов с использованием Википедии [21,8,18,10,11]. Для экспериментов, обсуждаемых в данной работе, был реализован метод, предложенный Д. Турдаковым и П. Велиховым в работе [21]. В [21] авторы используют страницы для многозначных терминов и перенаправляющие страницы Википедии. С использованием таких страниц Википедии строится набор возможных значений термина. Далее наиболее подходящее значение выбирается при помощи знаний о семантической близости терминов: для каждого возможного значения термина вычисляется степень его семантической близости с контекстом. В итоге выбирается то значение термина, степень семантической близости с контекстом которого было наибольшим.

Распространенной проблемой традиционных методов извлечения ключевых терминов является наличие абсурдных фраз в результате, таких как, например, "using", "electric cars are". Использование Википедии как контролирующего

тезауруса позволяет нам избежать данной проблемы: все ключевые термины, полученные в результате работы нашего метода, являются осмысленными фразами.

Результатом работы данного шага является список терминов, в котором каждый термин соотнесен с одной соответствующей статьей Википедии, описывающей его значение.

3.3. Построение семантического графа

На данном шаге по списку терминов, полученном на предыдущем шаге, мы строим семантический граф.

Семантический граф представляет собой взвешенный граф, вершинами которого являются термины документа, наличие ребра между двумя вершинами означает тот факт, что термины семантически связаны между собой, вес ребра является численным значением семантической близости двух терминов, которые соединяет данное ребро.

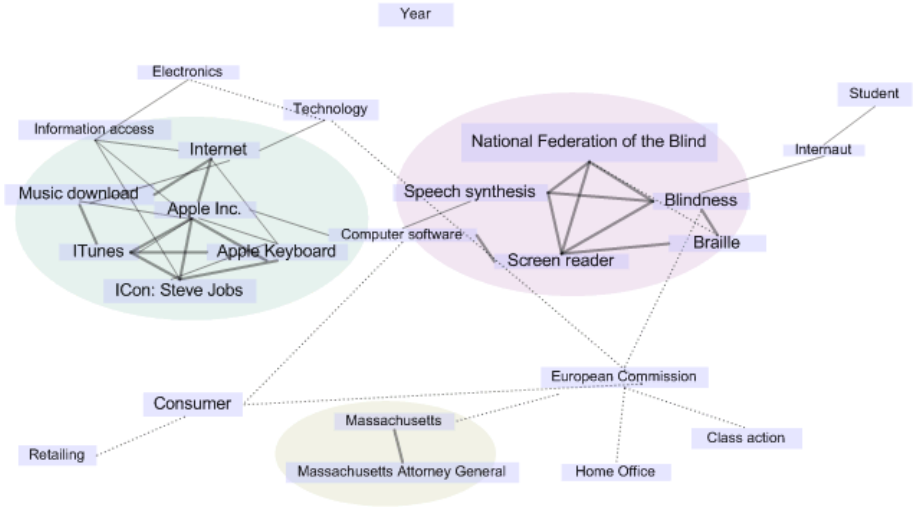

Рис. 1. Пример семантического графа, построенного по новостной статье «Apple to Make ITunes More Accessible For the Blind»

Рис. 1 демонстрирует пример семантического графа, построенного из новостной статьи *«Apple to Make ITunes More Accessible For the Blind»*. В статье говорится о том, что главный прокурор штата Массачусетс и Национальная Федерация Слепых достигли соглашения с корпорацией Apple Inc., следуя которому Apple сделает доступным свой музыкальный Интернет-сервис ITunes для слепых пользователей посредством технологии screen-reading. На рис. 1 можно видеть, что термины, релевантные *Apple Inc.* и

Blindness, образуют два доминантных сообщества, а термины *Student*, *Retailing* и *Year* оказались на периферии и слабо связаны с остальным графом.

Важно отметить тот факт, что термины – ошибки, возникшие при разрешении лексической многозначности терминов, проведенного на втором шаге, оказываются периферийными или даже изолированными вершинами графа, и не примыкают к доминантным сообществам.

3.4. Обнаружение сообществ в семантическом графе

Целью данного шага является автоматическое обнаружение сообществ в построенном семантическом графе. Для решения этой задачи мы применяем алгоритм Гирвана-Ньюмана. В результате работы алгоритма исходный граф разбивается на подграфы, которые представляют собой тематические сообщества терминов.

Для оценки качества разбиения некоторого графа на сообщества авторы [15] предложили использовать *меру модулярности* (*modularity*) графа. Модулярность графа является свойством графа и некоторого его разбиения на подграфы. Она является мерой того, насколько данное разбиение качественно в том смысле, что существует много ребер, лежащих **внутри** сообществ, и мало ребер, лежащих **вне** сообществ (соединяющих сообщества между собой). На практике значение модулярности, лежащее в диапазоне от 0.3 до 0.7, означает, что сеть имеет вполне различимую структуру с сообществами, и применение алгоритма обнаружения сообществ имеет смысл.

Мы отметили, что семантические графы, построенные из текстовых документов (таких как, например, новостная статья, или научная статья), имеют значение модулярности от 0.3 до 0.5.

3.5. Выбор подходящих сообществ

На данном шаге из всех сообществ необходимо выбрать те, которые содержат ключевые термины. Мы ранжируем все сообщества таким образом, чтобы сообщества с высокими рангами содержали важные термины (ключевые термины), а сообщества с низкими рангами – незначимые термины, а также ошибки разрешения лексической многозначности терминов, которые могут возникнуть на втором шаге работы нашего метода.

Ранжирование основано на использовании *плотности* и *информативности* сообщества. Плотностью сообщества является сумма весов ребер, соединяющих вершины этого сообщества.

Экспериментируя с традиционными подходами, мы обнаружили, что использование меры tf.idf терминов помогает улучшить ранжирование сообществ. Tf.idf дает большие коэффициенты терминам, соответствующим именованным сущностям (например, *Apple Inc.*, *Steve Jobs*, *Braille*), а терминам, соответствующим общим понятиям (таким как, например, *Consumer*, *Year*, *Student*) дает низкие коэффициенты. Мы считаем tf.idf для терминов, используя Википедию так, как описано в работе [8]. Под

информативностью сообщества мы понимаем сумму tf.idf-терминов, входящих в это сообщество, деленную на количество терминов сообщества.

В итоге, мы считаем ранг сообщества, как плотность сообщества, умноженная на его информативность, и сортирует сообщества по убыванию их рангов.

Приложение, использующее наш метод для извлечения ключевых слов, может использовать любое количество сообществ с наивысшими рангами, однако, на практике имеет смысл использовать 1-3 сообщества с наивысшими рангами.

4. Экспериментальная оценка

В этом разделе мы обсудим экспериментальные оценки предложенного метода. Поскольку не существуют стандартных бенчмарков для измерения качества извлеченных из текстов ключевых терминов, мы провели эксперименты с привлечением ручного труда, то есть **полнота** и **точность** извлеченных ключевых слов оценивались людьми – участниками эксперимента.

Мы собрали 30 блог-постов из следующих блогов технической тематики: «*Geeking with Greg*», автор Грег Линден, *DBMS2*, автор Курт Монаш, *Stanford Infoblog*, авторы – члены группы Stanford Infolab. В эксперименте приняли участие пять человек из отдела информационных систем ИСП РАН. Каждый участник должен был прочитать каждый блог-пост и выбрать в нем от 5 до 10 ключевых терминов. Каждый ключевой термин должен присутствовать в блог-посте, и для него должно быть найдено соответствующая статья в Википедии. Участники также были проинструктированы выбирать ключевые слова так, чтобы они покрывали все основные темы блог-поста. В итоге для каждого блог-поста мы выбрали такие ключевые термины, которые были выделены, по крайней мере, двумя участниками эксперимента. Названия перенаправляющих статей Википедии и название статей, на которые идет перенаправление, по сути, представляют собой синонимы, и мы в нашем эксперименте считали их **одним термином**.

Метод, представленный в данной статье, был реализован по следующим архитектурным принципам. Для достижения лучшей производительности мы не вычисляли семантическую близость всех пар терминов Википедии заранее. Данные, необходимые для подсчета семантической близости терминов на лету, а именно, заголовки статей Википедии, информация о ссылках между статьями, статистическая информация о терминах были загружены в оперативную память. В итоге полученная база знаний занимала в оперативной памяти 4.5 Гбайта. База знаний была установлена на выделенном компьютере с размером оперативной памяти, равным 8 Гбайт. Клиентское приложение работали с базой знаний посредством вызовов удаленных процедур.

4.1. Оценка полноты выделенных ключевых терминов

Под полнотой мы понимаем долю ключевых слов, выделенных вручную, которые так же были выделены автоматически нашим методом:

$$\frac{|\{\text{manually extracted}\} \bigcap \{\text{automatically extracted}\}|}{|\{\text{manually extracted}\}|}$$

где под {manually extracted} мы понимаем множество ключевых слов, извлеченных вручную участниками эксперимента для некоторого документа, под {automatically extracted} мы понимаем множество всех ключевых терминов, автоматически извлеченных нашим методом для того же документа. Знаком |S| мы обозначаем мощность множества S, то есть количество терминов в множестве S.

Для 30 блог-постов мы имеем 180 ключевых терминов, выделенных участниками эксперимента вручную, 297 – выделенных автоматически, 127 вручную выделенных ключевых слов были также выделены автоматически. Таким образом, полнота равно **68%**.

4.2. Оценка точности выделенных ключевых терминов

Мы оцениваем точность, используя ту же методологию, которой пользовались для оценки полноты. Под точностью мы понимаем долю тех слов, автоматически выделенных нашим методом, которые также были выделены вручную участниками эксперимента:

$$\frac{|\{\text{manually extracted}\} \bigcap \{\text{automatically extracted}\}|}{|\{\text{automatically extracted}\}|}$$

Итого, следуя показателям нашей тестовой коллекции, точность равна **41%**.

4.3. Пересмотр оценки полноты и точности

С целью более качественной оценки работы метода мы также пересмотрели наши оценки полноты и точности. Важной особенностью нашего метода является факт, что он в среднем выделяет **больше** ключевых слов, чем человек. Более точно, наш метод обычно извлекает больше ключевых терминов, релевантных одной теме. Например, рассмотрим рис. 1. Для темы, относящейся к *Apple Inc.*, наш метод выделяет термины: *Internet*, *Information access*, *Music download*, *Apple Inc.*, *ITunes*, *Apple Keyboard* и *Steve Jobs*, в то время как человек обычно выделяет меньше терминов и склонен выделять имена и названия: *Music download*, *Apple Inc.*, *ITunes* и *Steve Jobs*. Это означает, что, иногда метод извлекает ключевые термины с лучшим покрытием основных тем документа, чем это делает человек. Этот факт побудил нас пересмотреть оценку полноты и точности работы нашего метода.

Каждый участник эксперимента был проинструктирован пересмотреть ключевые термины, которые он сам выделил следующим образом. Для каждого блог-поста он должен был изучить ключевые термины, выделенные автоматически, и, по возможности, расширить свой набор ключевых слов, то есть дополнить его теми терминами, которые, по его мнению, относятся к главным темам документа, но не были выделены на первом этапе.

После такого пересмотра, мы получили 213 ключевых слов, выделенных вручную (вместо 180), таким образом, участники эксперимента добавили 33 новых ключевых термина, что означает, что наше предположение имело смысл, и такой пересмотр важен для полноценной оценки работы метода. В итоге, полнота равна **73%** и точность – **52%**.

5. Заключение

Мы предложили новый метод для извлечения ключевых терминов из текстовых документов. Одним из преимуществ нашего метода является отсутствие необходимости в предварительном обучении, поскольку метод работает над базой знаний, построенной из Википедии. Важной особенностью нашего метода является форма, в которой он выдает результат: ключевые термины, полученные из документа, сгруппированы по темам этого документа. Сгруппированные по темам ключевые термины могут значительно облегчить дальнейшую категоризацию данного документа и выведение его общей темы.

Эксперименты, проведенные с использованием ручного труда, показали, что наш метод позволяет извлекать ключевые термины с точностью и полнотой, сравнимой с теми, что дают современные существующие методы.

Мы отметили, что наш метод может быть с успехом применен для очистки сложных составных документов от неважной информации, и выделения главной темы в них. Это означает, что его интересно было бы применить для выделения ключевых терминов из Web-страниц, которые, как правило, загружены второстепенной информацией, например, меню, навигационные элементы, реклама. Это направление дальнейшей работы.

Список литературы

[1] Clauset, A.; Newman, M. E. J.; and Moore, C. 2004. Finding community structure in very large networks. Physical Review E 70:066111.

[2] Dakka, W., and Ipeirotis, P. G. 2008. Automatic extraction of useful facet hierarchies from text databases. In ICDE, 466–475. IEEE.

[3] de Solla Price, D. J. 1965. Networks of scientific papers. Science 169:510–515.

[4] Gabrilovich, E., and Markovitch, S. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In Proceedings of The Twentieth International Joint Conference for Artificial Intelligence, 1606–1611.

[5] Janik, M., and Kochut, K. J. 2008. Wikipedia in action: Ontological knowledge in text categorization. International Conference on Semantic Computing, 268–275.

[6] Kauffman, S. A. 1969. Metabolic stability and epigenesist in randomly constructed genetic nets. J Theor Biol 22(3):437–467.
[7] Manning, C. D., and Schtze, H. 1999. Foundations of Statistical Natural Language Processing. The MIT Press.
[8] Medelyan, O.; Witten, I. H.; and Milne, D. 2008. Topic indexing with wikipedia. In Wikipedia and AI workshop at the AAAI-08 Conference (WikiAI08).
[9] Mihalcea, R., and Csomai, A. 2007. Wikify!: linking documents to encyclopedic knowledge. In CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, 233–242. New York, NY, USA: ACM.
[10] Mihalcea, R. 2005. Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling. In HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing, 411–418. Morristown, NJ, USA: Association for Computational Linguistics.
[11] Mihalcea, R. 2007. Using wikipedia for automatic word sense disambiguation.
[12] Miller, G. A.; Fellbaum, C.; Tengi, R.; Wakefield, P.; Langone, H.; and Haskell, B. R. Wordnet: a lexical database for the english language. http://wordnet.princeton.edu/.
[13] Milne, D., and Witten, I. 2008. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In Wikipedia and AI workshop at the AAAI-08 Conference (WikiAI08).
[14] Milne, D. 2007. Computing semantic relatedness using wikipedia link structure. In Proceedings of the New Zealand Computer Science Research Student Conference (NZCSRSC).
[15] Newman, M. E. J., and Girvan, M. 2004. Finding and evaluating community structure in networks. Physical Review E 69:026113.
[16] Redner, S. 1998. How popular is your paper? An empirical study of the citation distribution. The European Physical Journal B 4:131.
[17] Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. Inf. Process. Manage. 24(5):513–523.
[18] Sinha, R., and Mihalcea, R. 2007. Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In ICSC '07: Proceedings of the International Conference on Semantic Computing, 363–369. Washington, DC, USA: IEEE Computer Society.
[19] Strube, M., and Ponzetto, S. 2006. WikiRelate! Computing semantic relatedness using Wikipedia. In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06), 1419–1424.
[20] Syed, Z.; Finin, T.; and Joshi, A. 2008. Wikipedia as an Ontology for Describing Documents. In Proceedings of the Second International Conference on Weblogs and Social Media. AAAI Press.
[21] Turdakov, D., and Velikhov, P. 2008. Semantic relatedness metric for wikipedia concepts based on link analysis and its application to word sense disambiguation. In Colloquium on Databases and Information Systems (SYRCoDIS).
[22] Wasserman, S.; Faust, K.; and Iacobucci, D. 1994. Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences). Cambridge University Press.