Инструментальные средства оценки качества научно-технических документов[1]

С.В. Герасимов, Р.В. Курынин, И.В. Машечкин, М.И. Петровский, Д.В. Царёв, А.А. Шестимеров
gerasimov@mlab.cs.msu.su, romaha@mlab.cs.msu.su, mash@cs.msu.su, michael@cs.msu.su, tsarev@mlab.cs.msu.su, andy@mlab.cs.msu.su
Факультет вычислительной математики и кибернетики
Московского государственного университета имени М.В.Ломоносова

Аннотация. В статье предлагается комбинированный подход к оценке качества научно-технических документов, учитывающий различные категории автоматически рассчитываемых характеристик качества документов — как существующие библиометрические и наукометрические характеристики (рассчитываемые на основе сведений из «цитатных» баз), так и новые типы характеристик, основанные на семантическом анализе текстов научно-технических документов, применении эвристических правил, а также на применении методов оценки наличия прямых текстовых заимствований (плагиата). На основе полученных базовых оценок формируется интегральный показатель оценки качества научно-технических документов с использованием методов машинного обучения аналогично решению задачи ранжирования в информационном поиске. Представлена разработанная экспериментальная система, основанная на предложенном подходе, а также приводятся экспериментальные исследования разработанной системы, направленные на проверку точности оценки научно-технических документов.

Проведённый в статье анализ состояния исследований в РФ и за рубежом в области методов оценки качества научно-технических документов показал, что предложенный в статье подход на основе автоматического расчета базовых оценок из указанного «расширенного» набора групп никем не рассматривалась в настолько широкой постановке и в целом является новаторским.

[1] Работа выполнена при поддержке государственного контракта №14.514.11.4016 и грантов РФФИ 11-07-00616, 12-07-00585.

Ключевые слова: оценка качества научно-технических документов; библиометрия; наукометрия; латентно-семантический анализ; неотрицательная матричная факторизация; тематическое моделирование; методы машинного обучения.

1. Введение

Исследования в области оценки качества научно-технических документов (научно-технических статей, диссертаций, отчетных материалов о НИОКР, заявочных документов на проведение НИОКР, патентной документации и др.) активно ведутся на протяжении более десяти последних лет [[1]]. Традиционно данную оценку осуществляют на основе информации о публикациях авторов и о цитировании их статей, которая собирается в «цитатных» базах [[2]]. Широкое распространение в мире получили цитатные базы, представленные компанией Thomson Reuters (бывший Institute for Scientific Information, ISI): Science Citation Index, Social Sciences Citation Index и Arts & Humanities Citation Index, а также Journal Citation Reports (JCR) [[3]]. Другим примером системы количественного определения значимости публикаций является система CiteSeer, разработанная в 1997 году [[4]]. В России примером «цитатной» базы служит «Российский индекс научного цитирования» (РИНЦ) [[5], [6]] — библиографическая база данных научных публикаций российских ученых.

В дальнейшем были показаны недостатки использования подобных систем, основанных на анализе цитируемости, для оценки качества научных документов [[7], [8]]. В настоящее время активно ведутся исследования по выработке новых индикаторов и методов оценки качества научно-технических документов. Так, в 2008–2011 гг. в рамках седьмой рамочной программы ЕС (European 7th Framework Programme) осуществлялся проект European Educational Research Quality Indicators (EERQI) [[9]], по итогам которого был разработан набор методик и индикаторов оценки качества научных документов на национальных европейских языках (английском, французском, немецком и шведском). Часть разработанных индикаторов использует наиболее популярные «цитатные» базы, а также были предложены новые техники вычисления индикаторов оценки качества [[10]]. Важно отметить, что разработанные в проекте EERQI характеристики, оценивающие «точность», «оригинальность» и «значимость», задаются экспертом, а не вычисляются автоматически. Так, например, характеристика «оригинальность» тесно связана с задачей поиска плагиата и может оцениваться автоматически.

Кроме того, согласно [[11], [12]] можно отметить характеристики, иллюстрирующие типичные ошибки авторов при написании статей: например, «слишком длинная аннотация к статье», «статья не содержит ключевых разделов», «слишком частое использование аббревиатур и/или незначимых слов». Очевидно, что эти характеристики можно легко автоматически вычислять путём задания экспертом соответствующих параметров. Данный

тип эвристических характеристик не встречается в существующих решениях по автоматической оценке качества документов.

Другим недостатком существующих подходов является отсутствие вычислительных критериев (характеристик, индикаторов) оценки качества документов, основанных на семантическом анализе текстового содержимого. Поэтому можно сделать вывод об актуальности исследования и разработки оценок качества научно-технических документов на естественных языках на основе анализа моделей семантики.

В статье предлагается комплексный подход к вычислению оценки качества анализируемых научно-технических документов, основанный на расчете следующих пяти групп базовых оценок: «семантические» (на основе семантических моделей документов); «ссылочные» (основанные на анализе цитирования документов); «репутационные» (использующие наукометрические оценки авторов и журналов, связанных с анализируемым документом); «оценки наличия плагиата» (оценка вероятности наличия прямых текстовых заимствований); «эвристические» (использующие заданные экспертом правила и словари). На основе полученных базовых оценок формируется интегральный показатель оценки качества документов с использованием методов машинного обучения аналогично подходу к решению задачи ранжирования в информационном поиске.

Настоящая статья имеет следующую структуру. В разделе 2 приводится аналитический обзор существующих подходов к оценке качества научно-технических документов. В разделе 3 подробно рассматриваются концепции предлагаемого подхода оценки качества научно-технических документов. Раздел 4 содержит сведения о разработанной экспериментальной системе, основанной на предложенном подходе. Раздел 5 посвящен экспериментальному исследованию разработанной системы.

2. Существующие подходы оценки качества научно-технических документов

Традиционный подход оценки качества научно-технических документов основывается на использовании библиометрических показателей, рассчитываемых на основе сведений из «цитатных» баз, содержащих расширенный набор сведений о публикациях (т.е. не только данные об авторах, заглавии, наименовании журнала, годе, томе, выпуске, страницах, но и о списке литературы, цитируемой и цитирующей данную статью) [[2]]. На основе этой информации вычисляются (по различным методикам) оценки качества как самих статей, так и их авторов и журналов публикаций. В качестве примеров наиболее известных в мире «цитатных» баз можно привести системы продукции компании Thomson Scientific (Science Citation

Index, Social Sciences Citation Index и Arts & Humanities Citation Index, Journal Citation Reports) [[3]], а также системы CiteSeer [[4]] и Google Scholar [[13]].

Поскольку в подавляющем большинстве случаев иностранные индексы («цитатные» базы) работают с англоязычными публикациями, определение библиографических показателей крайне затруднительно для неанглоязычных стран (таких, как Россия, Китай, Франция, Португалия, Япония и др.). Для решения указанной проблемы необходимо создание «цитатной» базы, содержащей необходимые сведения о статьях, публикующихся на национальных языках. В частности, в РФ в 2005 г. проходили исследования, посвященные разработке Российского индекса научного цитирования (РИНЦ) [[5], [6]].

В дальнейшем были показаны и другие недостатки использования подходов на основе анализа цитируемости для оценки качества научных документов, наиболее часто в литературе выделяют следующие [[7], [8]]:

- «шум» в оценке цитируемости, получаемый за счёт цитирования автором своих же статей (самоцитирование);
- «цитатные» базы сами по себе не включают все существующие статьи (проблема неполноты наполнения баз), кроме того, их наполнения сильно различаются друг от друга (различные журналы входят в различные «цитатные» базы), что приводит к различным оценкам одного и того же документа или автора.

В 2008–2011 гг. в рамках седьмой рамочной программы ЕС (European 7th Framework Programme) осуществлялся проект EERQI [[9]], направленный как на решение указанных проблем, так и на выработку новых подходов к оценке качества научно-технических документов. По итогам проекта был разработан набор методик и индикаторов оценки качества научных документов на национальных европейских языках (в качестве тестового набора использовались научные документы на английском, французском, немецком и шведском языках). Часть разработанных индикаторов по-прежнему использует наиболее популярные «цитатные» базы, но также были предложены новые техники вычисления индикаторов оценки качества. Разработанные индикаторы можно условно разделить на два типа:

- «Внутренние» (вычисляются лишь на основе текстового содержимого анализируемого документа) — расчет ведется на основе опросника, который заполняют эксперты. Используются три группы характеристик:
  - точность (англ. *rigour*) — оцениваются такие характеристики, как четкость изложения подхода, методов и результатов, а также полнота и корректность полученных результатов;
  - оригинальность (англ. *originality*) — оценивается новизна предлагаемых методов и подходов;

- значимость (англ. *significance*) — оценивается научный вклад в исследуемую проблемную область.

– «Внешние» (вычисляются с применением метаданных документа, т.е информации о документах из внешних по отношению к ним источников) — программно-инструментальный набор по определению следующих характеристик:
  - Библиографические (с использованием Google Scholar, Google Web Search и MetaGer):
    - количество опубликованных статей (для каждого автора),
    - количество цитирований (для каждого автора),
    - диапазон дат извлеченных публикаций,
    - количество цитирований в год,
    - количество цитирований каждой статьи,
    - g-индекс (улучшенная модификация h-индекса),
    - сопоставление авторства (на основе информации из Google Web Search и MetaGer).
  - На основе сервисов следующих социальных сетей Интернет (оценивается сопоставление авторства и заголовков статей): citeulike (www.citeulike.org), LibraryThing (www.librarything.com), Connotea (www.connotea.org), Mendeley (www.mendeley.com).

Не смотря на то, что в проекте EERQI была предложена идея вычисления индикаторов на основе содержимого анализируемого документа («внутренний» тип индикаторов), все указанные характеристики данного типа, оценивающие точность, оригинальность и значимость, задаются экспертом, а не вычисляются автоматически. Хотя, например, характеристика «оригинальность» тесно связана с задачей поиска плагиата и может оцениваться автоматически. Кроме того, никак не используется семантический анализ текстового содержимого для вычисления каких-либо оценок качества документов.

По итогам обзора публикаций [[11], [12]], посвящённых экспертной оценке качества научных статей, можно отметить характеристики, иллюстрирующие типичные ошибки авторов при написании статей, например, «слишком длинная аннотация к статье», «статья не содержит ключевых разделов», «слишком частое использование аббревиатур и/или незначимых слов». Очевидно, что подобные характеристики можно автоматически вычислять путём задания экспертом соответствующих параметров — рекомендуемой длины аннотации, названия разделов, частоты использования аббревиатур, списка стоп-слов и частоты их встречаемости. Данный тип эвристических характеристик также не встречается в существующих решениях по автоматической оценке качества документов.

3. Концепции предлагаемого подхода оценки качества научно-технических документов

В рамках данной работы авторами предлагается комплексный подход автоматического вычисления оценки качества научно-технических документов, основанный на расчете следующих групп базовых оценок:

– «семантические» — оценки, рассчитываемые с использованием семантического анализа текстового содержимого документов;

– «ссылочные» — библиометрические оценки, основанные на анализе цитирования документов;

– «репутационные» — оценки, использующие библиометрическую и наукометрическую информацию об авторах и журналах документов из внешних по отношению к ним источников (метаданные документов);

– «оценка наличия плагиата» — оценка вероятности наличия прямых текстовых заимствований;

– «эвристические» — оценки, использующие заданные экспертом правила и словари.

При анализе каждого научно-технического документа вычисляются значения каждой базовой оценки, и на их основе рассчитывается интегральный показатель оценки качества с использованием методов машинного обучения аналогично решению задачи ранжирования в информационном поиске. В предлагаемом подходе проводится расчет всех групп базовых оценок и интегрального показателя как для исходного анализируемого документа, так и для коллекции семантически близких ему документов, формируемой в процессе анализа исходного документа.

Ниже в настоящем разделе приводятся описания подходов, используемых для автоматического вычисления всех групп базовых оценок и для расчёта результирующего интегрального показателя.

3.1 «Семантические» базовые оценки

Семантические оценки в предлагаемом подходе основаны на анализе тематических моделей семантики отдельных документов и коллекций документов [[14], [15], [16], [17]]. Тематические модели объединяют семантически схожие термы в тематики, при этом выделенные тематики ставятся в соответствие текстовым фрагментам. При тематическом моделировании отдельного документа в качестве текстовых фрагментов использовались предложения текста, при тематическом моделировании коллекции документов — сами документы.

Для тематического моделирования используется метод латентно-семантического анализа (англ. *Latent semantic analysis, LSA*) [[14], [15], [16]]. Основная идея данного метода состоит в том, что совокупность всех текстовых фрагментов приводит к взаимным ограничениям использований термов, определяющим сходство семантических значений термов и фрагментов. Латентно-семантический анализ работает с векторным представлением типа *"bag-of-words"* текстовых фрагментов [[14]]. Таким образом, анализируется числовая матрица отображения терм-фрагмент, строки которой соответствуют термам, а столбцы — фрагментам. Объединение термов в тематики и представление фрагментов в пространстве тематик осуществляется путём применения к данной матрице одного из матричных разложений. В настоящее время наиболее популярными матричными разложениями являются сингулярное разложение (англ. *Singular Value Decomposition, SVD*) [[14]] и неотрицательная матричная факторизации (англ. *Non-negative Matrix Factorization, NMF*) [[15]]. После применения к текстовой матрице одного из указанных матричных разложений формируется семантическая модель совокупности текстовых фрагментов, состоящая из:

1. Матрицы отображения пространства термов в пространство тематик.
2. Вектора, чьи элементы соответствуют весу выделенных тематик в тексте (или диагональной матрицы, чьи диагональные элементы соответствуют весам тематик). Для сингулярного разложения веса выделенных тематик аппроксимируются квадратом соответствующих сингулярных чисел [[18], [19]]. В случае применения неотрицательной матричной факторизации веса тематик предлагается рассчитывать на основе оригинального метода, разработанного коллективом авторов, для которого была показана применимость для задачи выделения ключевых предложений текста (автоматического аннотирования) [[20], [21], [22], [23]].
3. Матрицы представления текстовых фрагментов в пространстве тематик.

На основе получаемой информации о тематиках документов и коллекций документов вычисляются следующие семантические оценки качества документов:

1. *Оценки информационной сжимаемости документа (тематическое моделирование отдельного документа).* Используя веса выделенных тематик в качестве оценок количества информации, содержащейся в них, можно определить минимальное число тематик, требующееся для покрытия заданного процента информации, содержащейся в тексте[19]. Аналогично, используя тематическое представление текста документа, можно вычислить значимость (релевантность) каждого предложения текста [[20]], которую также можно использовать в качестве оценки количества информации, содержащейся в них. Таким образом, задав сетку информационных процентов от 10% до 100% с шагом 10%, получаем списки, состоящие из 10 элементов, содержащие спектр оценок информационной сжимаемости документа по его предложениям и тематикам, соответственно, для сингулярного разложения и неотрицательной матричной факторизации.
2. *Оценки принадлежности документа к набору тематик коллекции документов (тематическое моделирование коллекции документов).* Одним из результатов построения любой из рассматриваемых семантических моделей для коллекции семантически близких документов является представление документов в пространстве тематик коллекции. Таким образом, в качестве базовых семантических оценок качества документа использовался вектор его представления в пространстве тематик коллекции, получаемый при тематическом моделировании коллекции с помощью матричной факторизации [[15], [24]].

3.2 «Ссылочные» и «репутационные» базовые оценки

Предлагается в расчетах «ссылочных» и «репутационных» оценок использовать значения характеристик, приведенных в табл. 1 (для характеристик, значения которых получить не удается, допускается использовать значение «неизвестно»). Данные характеристики определяются для всех документов, участвующих в расчетах «ссылочных» и «репутационных» оценок [[6]].

Табл. 1 Характеристики ссылочных и репутационных оценок.

Категория	Характеристика	Способ получения
Информация о документе	год издания	на основе метаданных
	длина списка библиографии	на основе метаданных
	количество цитирований	на основе метаданных, из информационной базы или из внешних источников
	значение PageRank	рассчитываемая величина
	информация о распределении значений PageRank цитируемых	рассчитываемая величина

Категория	Характеристика	Способ получения
	документов	
	информация о распределении длин библиографических списков цитируемых статей	рассчитываемая величина
	информация о распределении значений индексов Хирша авторов документа	рассчитываемая величина
Информация о каждом авторе документа	индекс Хирша	из информационной базы или из внешних источников
Информация об организациях, в которых состоят авторы	индекс Хирша	из информационной базы или из внешних источников
Информация о журнале, в котором публикуется документ	классический импакт-фактор	из информационной базы или из внешних источников

В данной таблице под *значением PageRank* понимается рассчитанное значение, получаемое на основе применения алгоритма PageRank, описанного в [[25]], к некоторому набору документов, в качестве связей между которыми выступает цитирование (т.е. объектом анализа является ориентированный граф цитирования, где узлы соответствуют документам, а дуги — ссылкам одних документов на другие).

В качестве *информации о распределении значений* какой-либо величины понимается совокупность рассчитанных значений следующих статистических характеристик:

- минимум;
- первый квартиль (известный в литературе также, как нижний квартиль, или 0,25-квантиль);
- медиана (0,5-квантиль);
- третий квартиль (или верхний квартиль, или 0,75-квантиль);
- максимум.

В третьей колонке таблицы приводятся способы определения значения характеристики:

- на основе метаданных — значение характеристики определяется на основе текстового содержимого самого документа;
- из информационной базы — значение характеристики определяется из информационной базы, представляющей собой локальное хранилище, содержащее наукометрические и библиометрические данные о документах, полученные из внешних «цитатных» баз;
- из внешних источников — значение определяется с использованием данных из внешних источников (например, из какой-либо «цитатной» базы);
- рассчитываемая величина — значение характеристики рассчитывается в ходе проведения анализа по расчету «ссылочных» и «репутационных» оценок.

3.3 «Оценка наличия плагиата»

Для анализа документов на наличие прямых текстовых заимствований предлагается подход, основанный на следующих этапах:

- Разбивка исходного документа на текстовые фрагменты (англ. *chunking*).
- Получение коллекции документов для анализа на основе поискового запроса, сформированного из ключевых слов анализируемого документа (выделение ключевых слов происходит на стадии семантического анализа текста [[29]]), и библиографии документа.
- Фильтрация полученного набора для выделения подмножества документов для детального сравнения. На основе сравнения алгоритмов поиска заимствований на коллекции PAN [[26]] для фильтрации документов предлагается использовать представление в виде символьных n-грамм и n-грамм из слов. В качестве меры сходства предлагается использовать меру Жаккара [[27]] и скалярное произведение.
- Детальный анализ схожести двух документов. При определении численной характеристики присутствия заимствований предлагается использовать метод *String Sequence Kernel (SSK*) [[28]], устойчивый к локальным модификациям заимствованного текста. При использовании метода предлагается следующая модификация: в суммарную оценку сходства двух документов входит сходство для длин подпоследовательностей на отрезке [k, n] (где k и n — минимальная и максимальная длины подпоследовательностей), что

уменьшает влияние совпадающих устойчивых слов или выражений, связанных с общей тематикой сравниваемых текстов.

3.4 «Эвристические» оценки текста

В задачах поиска плагиата и Web-поиска часто используются статистические характеристики — эвристики, направленные на оценку «читабельности» (англ. *readability*) текста, т.е. простоту восприятия текста, и на оценку его стилистического написания [[30], [31], [32]]. К примерам эвристик «читабельности» можно отнести такие характеристики, как средняя длина предложений (в словах) и средняя длина слов (в символах) текста [[32]]. В качестве стилистических эвристик можно привести количество употребления частиц, предлогов, а также среднюю длину предложений и слов [[30], [31]]. В работе [[30]] показывается, как анализ набора стилистических характеристик позволяет определить принадлежность текста к конкретному автору.

Как видно даже из приведённых примеров, наборы эвристик для оценок «читабельности» и стилистического написания пересекаются. Поэтому приведём единый список наиболее популярных эвристик рассматриваемых категорий, которые используются в качестве эвристических базовых оценок качества научно-технических документов:

1. Среднее количество слов в предложениях (средняя длина предложений);
2. Среднее количество символов в словах (средняя длина слов);
3. Доля слов длиннее 7 символов;
4. Частота употребления стоп-слов (предлогов, союзов, частиц) в тексте;
5. Средняя по предложениям частота употребления стоп-слов;
6. Среднее количество знаков пунктуации на предложение;
7. Количество знаков экспрессивной пунктуации («!», «?», «…»);
8. Среднее по предложениям количество знаков экспрессивной пунктуации.

3.5 Расчёт интегрального показателя оценки качества научно-технических документов

В рамках данной работы предлагается подход вычисления оценки качества научно-технического документа как интегральной оценки на основе комбинации индивидуальных базовых оценок качества — семантических оценок, репутационных, ссылочных, эвристических, а также оценок уровня текстуального заимствования. Данный подход основан на использовании методов машинного обучения аналогично подходу к решению задачи ранжирования в задачах информационного поиска. Алгоритмы машинного обучения обладают возможностью обобщения, они способны корректно ранжировать или классифицировать документы, которые не встречались непосредственно в обучающей выборке. В нашем случае предлагается процедура расчета интегральной оценки, включающей три этапа:

1. Формирование обучающей выборки документов — эксперт ранжирует документы из выборки по их качеству, т.е. задает относительный порядок на основе попарных сравнений. На основе заданных попарных сравнений качества документов строится оценка относительных рангов документов коллекции на шкале с использование модели Бредли-Терри с ничьей [[33]]. Оценки попарных сравнений качества документов могут задаваться экспертно или вычисляться на основе любой из базовых оценок (например, индекса цитируемости).
2. Строится модель ранжирования в виде функции экспоненциальной регрессии [[34]] с пошаговым выбором независимых переменных (для борьбы с переобучением) [[35]] для прогнозирования относительного ранга, рассчитанного на предыдущем этапе.
3. Построенная модель ранжирования используется для прогноза ранга научно-технического документа относительно заданной коллекции семантически близких ему документов. Относительный ранг в рамках коллекции семантически близких документов рассматривается как оценка качества научно-технического документа.

4. Экспериментальная система оценки качества научно-технических документов

Данный раздел статьи посвящён архитектуре экспериментальной системы (ЭС), выполняющей вычислительную оценку качества научно-технических документов. В разделе приводятся описания программных компонент, входящих в состав ЭС, и используемые технологии.

Реализация ЭС предполагает проведение многоэтапной разнородной обработки поданных на вход документов. Приведём описание предложенных этапов обработки коллекции документов в виде описаний задач, которые выполняются на том или ином этапе.

1. Этап «Базовая обработка» (БО) — выполняется с каждым анализируемым документом. Данный этап обработки документа включает в себя три задачи:
    - Первичная обработка: выделение из файла анализируемого документа текстового содержимого (текста) и метаинформации, включающей авторов и заголовок документа, список библиографии; идентификация документа по метаинформации в информационной базе системы.
    - Обработка текста: вычисление оценок информационной сжимаемости и эвристических оценок документа; выделение ключевых слов документа.
    - Обработка метаинформации: поиск библиометрической и наукометрической информации во внешних источниках; вычисление ссылочных и репутационных оценок качества документа.
2. Этап «Формирование контекста». Данный этап состоит из одной задачи: для каждого анализируемого документа выполняется поиск документов, сходных с анализируемым, по ключевым словам, сформированным на предыдущем этапе, в сети Интернет и локальном хранилище системы (представляющим собой индексированный массив научно-технических документов); объединение всех найденных документов с анализируемыми в расширенный контекст.
3. Этап «Базовая обработка контекста». Данный этап состоит из одной задачи — выполнение базовой обработки документов расширенного контекста, для которых данная обработка ещё не была произведена.
4. Этап «Семантический анализ контекста». Данный этап состоит из одной задачи: проведение вычисления оценки принадлежности документов к набору тематик расширенного контекста документов (тематическое моделирование коллекции документов), на основе которого производится построение семантического контекста — коллекции семантически близких документов, полученной путем удаления из расширенного контекста документов, семантически далеких от исходных анализируемых.
5. Этап «Оценка текстовых заимствований». Данный этап состоит из одной задачи — вычисление базовых оценок качества всех документов семантического контекста на основе оценок степени взаимного текстового заимствования.

6. Этап «Ранжирование». Данный этап обработки документов включает в себя две задачи:
    - формирование попарных сравнений качества документов;
    - ранжирование документов (вычисление интегрального показателя).

Часть задач, входящих в рассмотренные этапы обработки, может выполняться параллельно (например, обработка текста и обработка метаинформации в «Базовой обработке»); другие задачи связанны по входным/выходным данным и требуют последовательного исполнения (например, цепочка последовательного выполнения задач этапов «Формирование контекста», «Базовая обработка контекста», «Семантический анализ контекста»). При этом результаты отдельных задач могут быть использованы сразу в нескольких других задачах (например, исходный текст документа необходим при вычислении задачи обработки текста в «Базовой обработке» и задачи этапа «Оценка текстовых заимствований»). Некоторые этапы обработки могут занимать значительное время (например, обработка метаинформации в «Базовой обработке», т.к. в этой задаче выполняется анализ внешних источников наукометрической информации, вычисление ссылочных и репутационных базовых оценок). Поэтому важно, чтобы пересчет последующих этапов обработки (например, в случае аппаратного сбоя) не приводил к пересчету результатов нижележащих этапов.

Перечисленные особенности процедуры обработки исходных документов привели к необходимости создания в рамках ЭС подсистемы управления вычислительными задачами (Диспетчер задач). Таким образом, диспетчер задач должен удовлетворять следующим требованиям:

1. Разбиение исходной процедуры обработки научно-технических документов на этапы, предполагающие как последовательное, так и параллельное исполнение.
2. Возможность долговременного хранения результатов вычислений.
3. Возможность пересчета выбранных этапов и этапов, использующих их, без необходимости полного пересчета всех задач.
4. Распределенное вычисление задач.
5. Восстановление корректного состояния вычислений после аппаратного сбоя.
6. Мониторинг состояния вычислений.

На рис. 1 представлены программные компоненты, реализующие задачи этапов обработки, и программный компонент управления ими.

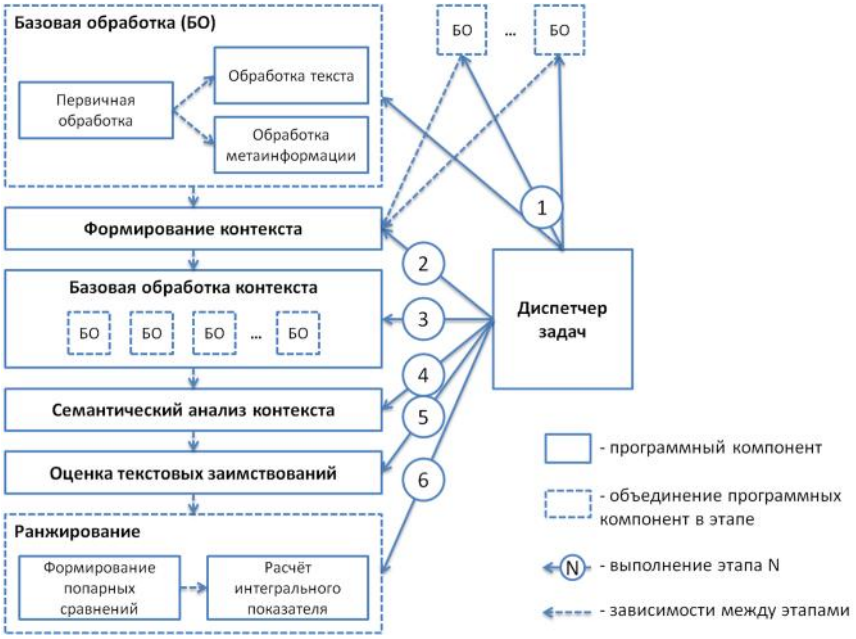

Рис. 1. *Программные компоненты ЭС.*

Диспетчер задач, реализующий механизм вычислительных задач ЭС, основывается на фреймворке Twisted Framework [[36]]. Использование асинхронного программирования позволило эффективно организовать в ЭС вычисления, обусловленные всеми стадиями обработки научно-технических документов, исполняя «легкие» операции и операции ввода-вывода в основной нити Twisted, направляя относительно ресурсоемкие вычисления в пул нитей, а вычисления, требующие значительного процессорного времени, выносить в отдельные исполняемые модули, написанные на языке программирования C++ и исполняемые в отдельных процессах ОС либо на отдельных физических машинах, связанных локальной сетью. Так, задача обработки метаинформации, вычисляющая репутационные и ссылочные оценки, реализована на языке Python и выполняется в основной нити Twisted, а задача расчета интегрального показателя реализована на языке C++ и выполняется в отдельном процессе ОС.

Указанный механизм задач обладает поддержкой регулирования степени параллелизма, используемого при вычислении значений задач. Механизм позволяет задать общее ограничение на число задач, параллельно исполняемых в рамках процессов ОС.

Исходные тексты ЭС и все используемые в ЭС сторонние библиотеки являются кросс-платформенными в рамках ОС семейств Unix (ядро Linux версии 2.4 и выше) и MS Windows.

Пользовательский интерфейс ЭС реализован на языке JavaScript в дополнительном программном компоненте взаимодействия с пользователем. Реализация базируется на использовании Ajax-фреймворка для создания пользовательского интерфейса «насыщенных» Интернет-приложений qooxdoo [[37]], поддерживающего следующие версии веб-браузеров: Internet Explorer 6+, Firefox 2+, Opera 9+, Safari 3.0+ и Chrome 2+.

5. Результаты экспериментального исследования

Настоящий раздел посвящен проведению экспериментальных исследований ЭС по проверке точности рассчитываемой оценки качества научно-технических документов на русском и английском языках.

Для проведения экспериментов были сформированы тестовые наборы данных, содержащие целевые документы:

- первый тестовый набор сформирован из русскоязычных публикаций. В него вошли статьи из конференции «Математические методы распознавания образов» 2011 г. [[38]] и тематического сборника факультета ВМК МГУ имени М.В.Ломоносова «Программные системы и инструменты» 2012 г.;
- второй тестовый набор сформирован из англоязычных публикаций. В него вошли статьи из конференций ICDM 2006 [[39]] и ICMET 2011 [[40]].

Каждый тестовый набор состоит из 10 документов, написанных на схожие тематики. При этом документы каждого тестового набора были размечены экспертом на две категории «А» и «Б» по 5 документов. Категоризация документов каждого тестового набора выполнялась по следующей эвристике: документ категории «Б» имеет худшее качество по сравнению с документом категории «А» для любых документов из тестового набора.

Для каждого тестового набора прошло по одной серии экспериментов. Каждая серия экспериментов заключалась в формировании различных правил разметки тестового набора на основе попарных сравнений, т.е. в задании правил вида «документ Д1 лучше документа Д2», и последующем ранжировании тестового набора. При этом сформированные правила попарных сравнений различались как по составу участвующих документов, так и по количеству самих правил. Т.к. каждый тестовый набор состоит из 5 документов категории «А» и 5 документов категории «Б», то всего можно сформировать 25 различных правил попарных сравнений (без ничьих).

В дальнейшем для оценки точности ранжирования использовался критерий *Ranking Loss* [[41]], который отражает среднюю долю некорректно

упорядоченных пар документов (рис. 2). Пара документов считается некорректно упорядоченной, если документ категории «Б» располагался выше документа категории «А» в соответствии с рассчитанным рангом. Чем меньше значение *Ranking Loss*, тем лучше качество ранжирования. Качество ранжирования совершенно, когда *Ranking Loss = 0*. Итоговая точность ранжирования оценивалась в процентах и вычислялась по формуле: 100*(1 – *Ranking Loss*).

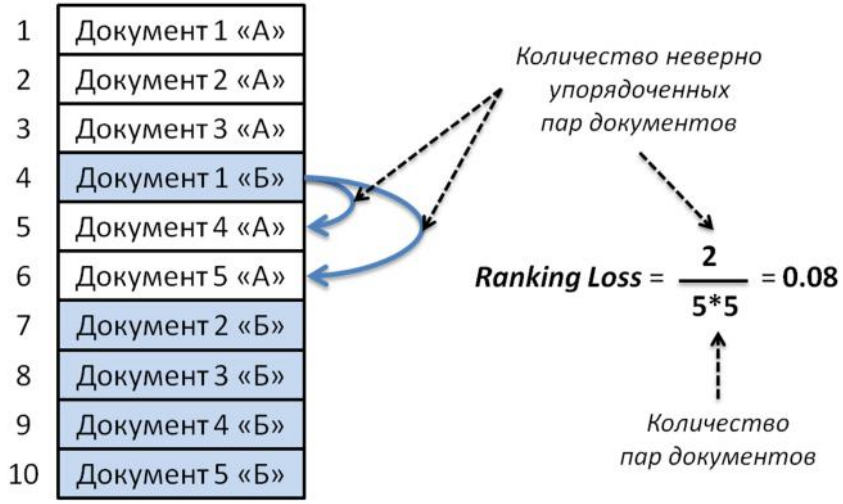

Рис. 2. *Пример вычисления критерия Ranking Loss для оценки качества ранжирования.*

В результате экспериментального исследования было получено, что достаточно сформировать 4-5 правил попарных сравнений, что составляет 16-20% от общего числа различных правил, для получения точности не меньшей 98.8%.

6. Заключение

В статье предлагается комплексный подход к оценке качества научно-технических документов, включающий автоматический расчет следующих групп базовых оценок качества: семантические, построенные на основе семантических моделей отдельных документов и коллекции семантически близких документов; библиометрические и наукометрические, основанные на анализе графа цитирования документов и использующие репутационные оценки авторов и изданий; оценки наличия прямых текстовых заимствований из семантически близких документов; эвристические, использующие заданные экспертом правила и словари. На основе полученных базовых оценок

формируется интегральный показатель оценки качества научно-технических документов с использованием методов машинного обучения аналогично решению задачи ранжирования в информационном поиске. Для этого на основе попарных сравнений качества документов коллекции, заданных экспертно или вычисленных на основе базовых оценок, строится модель ранжирования, которая определяет интегральную оценку как функцию от совокупности базовых оценок, представленных выше типов. Полученная функция ранжирования далее применяется для вычисления интегральной оценки качества документов, семантически близких к данной коллекции.

В статье были представлены экспериментальные исследования по проверке точности рассчитываемой оценки качества научно-технических документов на русском и английском языках. Для этого были разработаны тестовые наборы данных, содержащие целевые документы на русском и английском языках, соответственно. Исследования показали достаточность задания 4-5 правил попарных сравнений, что составляет 16-20% от общего числа различных правил, для получения точности оценки документов не меньшей 98.8%.

В настоящее время проблема оценки качества научно-технических документов на основе автоматического расчета базовых оценок из предложенного «расширенного» набора групп никем не решена, и даже не рассматривалась в настолько широкой постановке. На основе приведённого анализа состояния исследований в РФ и за рубежом можно сделать вывод, что предложенный в статье подход для решения данной проблемы в целом является новаторским.

7. Список литературы

[1]. Steve Lawrence, Kurt Bollacker, C . Lee Giles. Indexing and Retrieval of Scientific Literature // Eighth International Conference on Information and Knowledge Management, CIKM 99, Kansas City, Missouri, November 2–6, pp. 139–146, 1999.

[2]. В.В. Писляков. Методы оценки научного знания по показателям цитирования // М.: Социологический журнал, 2007, N1, стр. 128-140.

[3]. Официальный сайт ISI Web of Knowledge (ныне подразделение Healthcare & Science business в Thomson Reuters) // http://www.webofknowledge.com.

[4]. Официальный сайт системы CiteSeer // http://citeseerx.ist.psu.edu.

[5]. Российский Индекс Научного Цитирования // http://elibrary.ru/project_risc.asp.

[6]. Писляков В. В. Наукометрические методы и практики, рекомендуемые к применению в работе с российским индексом научного цитирования // Отчёт о научно-исследовательской работе (промежуточный) по теме «Разработка системы статистического анализа российской науки на основе данных российского индекса цитирования». — М., 2005.

[7]. Meho L (Meho, Lokman); Yang K (Yang, Kiduk). Fusion approach to citation-based quality assessment // Proceedings Of Issi 2007: 11th International Conference Of The International Society For Scientometrics And Informetrics, Vols I And II : 568-581.

[8]. Angela Vorndran, Alexander Botte. Analysis and evaluation of existing methods and indicators for quality assessment of scientific publications // http://www.eerqi.eu/sites/default/files/Analysis_and_evaluation_of_existing_methods_and_indicators.pdf [PDF].
[9]. Официальный сайт проекта EERQI – European Educational Research Quality Indicators // www.eerqi.eu.
[10]. EERQI Project Final Report (2011) // http://eerqi.eu/sites/default/files/Final_Report.pdf [PDF].
[11]. Moyses Szklo. Quality of scientific articles // Revista Saúde Pública vol.40 special issue São Paulo Aug. 2006.
[12]. Dr Navneet Gupta BSc (Hons) PhD MCOptom FBCLA. How to Evaluate a Scientific Research Article // http://www.optometry.co.uk/uploads/articles/ARTICLE%200309.pdf [PDF].
[13]. Официальный сайт системы Google Scholar// http://scholar.google.ru.
[14]. Berry M.W., Dumais S.T., O'Brien G.W. Using Linear Algebra for Intelligent Information Retrieval // University of Tennessee Knoxville. TN. USA, 1994.
[15]. Lee D.D., Seung H.S. Learning the parts of objects by non-negative matrix factorization // Nature, 401, pp. 788-791, 1999.
[16]. Rakesh P., Shivapratap G., Divya G., Soman KP. Evaluation of SVD and NMF Methods for Latent Semantic Analysis // International Journal of Recent Trends in Engineering, Vol. 1, No. 3, 2009.
[17]. Griffiths T L, Steyvers M. Finding scientific topics // In: Proceedings of the National Academy of Sciences. USA, 2004, 101: 5228–5235.
[18]. Steinberger J., Ježek K. Text Summarization and Singular Value Decomposition // In Lecture Notes for Computer Science vol. 2457, Springer-Verlag, 2004, pp. 245-254.
[19]. Steinberger J. Text Summarization within the LSA Framework // PhD Thesis, University of West Bohemia in Pilsen, Czech Republic, January 2007.
[20]. Машечкин И.В., Петровский М.И., Царёв Д.В. Методы вычисления релевантности фрагментов текста на основе тематических моделей в задаче автоматического аннотирования // Вычислительные методы и программирование. Том 14, 2013. 91-102.
[21]. Mashechkin I.V., Petrovskiy M.I., Popov D.S., Tsarev D.V. Automatic text summarization using latent semantic analysis // Programming and Computer Software, pp. 299-305, 2011.
[22]. Tsarev D., Petrovskiy M., Mashechkin I. Using NMF-based text summarization to improve supervised and unsupervised classification // 11th International Conference on Hybrid Intelligent Systems (HIS), Malacca, MALAYSIA. P. 185-189, 2011.
[23]. Dmitry Tsarev, Mikhail Petrovskiy and Igor Mashechkin, Supervised and Unsupervised Text Classification via Generic Summarization International Journal of Computer Information Systems and Industrial Management Applications. MIR Labs, Volume 5, 2013, pp. 509-515.
[24]. Wei Xu, Xin Liu, Yihong Gong Document clustering based on non-negative matrix factorization // Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, Toronto, Canada, 2003.
[25]. Y. Ding. Applying weighted PageRank to author citation networks. In Proceedings of JASIST. 2011, pp. 236-245.

[26]. M. Potthast, T. Gollub, M. Hagen, J. Graßegger, J. Kiesel, M. Michel, A. Oberländer, M. Tippmann, A. Barrón-Cedeño, P. Gupta, P. Rosso, B. Stein. Overview of the 4th International Competition on Plagiarism Detection. CLEF2012. 2012.
[27]. S. Alzahrani, N. Salim. Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection, Lab Report for PAN at CLEF2010, 2010.
[28]. A. Martins. String kernels and similarity measures for information retrieval. 2006.
[29]. Berry M.W., Browne M., Langville A.N., Pauca V.P., Plemmons R.J. Algorithms and applications for approximate nonnegative matrix factorization // Computational Statistics and Data Analysis, pp. 155-173, 2007.
[30]. Фоменко В.П., Фоменко Т.Г. Авторский инвариант русских литературных текстов. Предисловие А.Т. Фоменко. // Фоменко А.Т. Новая хронология Греции: Античность в средневековье. Т. 2. М.: Изд-во МГУ, 1996, с.768-820.
[31]. Braslavski P. Document Style Recognition Using Shallow Statistical Analysis. In Proceedings of the ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP, Nancy, France, 2004, p. 1–9.
[32]. DuBay, W.H. The Principles of Readability. Costa Mesa, CA: Impact Information. 2004.
[33]. P.V. Rao and L.L. Kupper, "Ties in paired-comparison experiments: A generalization of the Bradley–Terry model", Amer. Statist. Assoc, 62, 1967, pp. 194–204.
[34]. Turner, H and Firth, D (2012). Bradley-Terry Models in R: The BradleyTerry2 Package. Journal of Statistical Software 48(9), 1–21.
[35]. Hastie, Tibshirani and Friedman (2008). The Elements of Statistical Learning (2nd edition) Springer-Verlag. 763 pages.
[36]. Официальный сайт Twisted Framework // http://twistedmatrix.com.
[37]. Официальный сайт qooxdoo // http://qooxdoo.org.
[38]. Конференция «Математические методы распознавания образов» // http://www.mmro.ru.
[39]. The IEEE International Conference on Data Mining (ICDM) // http://www.cs.uvm.edu/~icdm.
[40]. International Conference on Mechanical and Electrical Technology (ICMET) // http://www.icmet.ac.cn.
[41]. Zhang M.-L., Zhou Z.-H. A k-nearest neighbor based algorithm for multi-label classification // Proceedings of the 1st IEEE International Conference on Granular Computing (GrC'05). Beijing, China, 2005. pp. 718-721.

Tools for Quality Assessment of Scientific and Technical Documents

S.V. Gerasimov, R.V. Kurynin, I.V. Mashechkin, M.I. Petrovskiy, Tsarev D.V., A.A.Shestimerov
Moscow State University, Moscow, Russia
gerasimov@mlab.cs.msu.su, romaha@mlab.cs.msu.su, mash@cs.msu.su, michael@cs.msu.su, tsarev@mlab.cs.msu.su, andy@mlab.cs.msu.su

Abstract: In the paper the complex approach to scientific and technical document quality assessment is proposed based on various automatically calculated document quality characteristics as widely used bibliometric and scientometric (based on citation indices), and the new types of characteristics based on the text semantic analysis, heuristics, and also on plagiarism detection methods. The integrated indicator of scientific and technical document quality assessment is formed on the basis of the received basic characteristics with use of machine learning methods similar to the problem of ranking in information retrieval. The developed prototype system based on offered approach is presented, and also the experimental investigations of the developed system directed on check of scientific and technical document quality assessment accuracy are carried out. The analysis of the state of art researches of scientific and technical document quality assessment showed the offered approach based on enhanced list of basic characteristic groups was considered by nobody in so broad statement and as a whole is innovative. The main part of the paper has the following structure. The second section contains an analytical overview of existing approaches to assess quality of scientific and technical documents. The third section provides detail of a proposed approach to assess quality of scientific and technical documents. The forth section describes a prototype system based on the proposed approach. The fifth section discusses results of experiments.

Keywords: scientific and technical document quality assessment; bibliometrics; scientometrics; latent semantic analysis; non-negative matrix factorization; topic model; machine learning

References

[1]. Steve Lawrence, Kurt Bollacker, C . Lee Giles. Indexing and Retrieval of Scientific Literature. Eighth International Conference on Information and Knowledge Management, CIKM 99, Kansas City, Missouri, November 2–6, pp. 139–146, 1999.
[2]. V.V. Pisljakov. Metody ocenki nauchnogo znanija po pokazateljam citirovanija [Methods of assessment of scientific knowledge in terms of citation]. M.: Sociologicheskij zhurnal [Joutnal of Sociology], 2007, N1, str. 128-140 (in Russian).
[3]. ISI Web of Knowledge. http://www.webofknowledge.com.
[4]. CiteSeer. http://citeseerx.ist.psu.edu.
[5]. Rossijskij Indeks Nauchnogo Citirovanija [Russian Science Citation Index]. http://elibrary.ru/project_risc.asp (in Russian).
[6]. Pisljakov V. V. Naukometricheskie metody i praktiki, rekomenduemye k primeneniju v rabote s rossijskim indeksom nauchnogo citirovanija [Scientometric methods and practices that are recommended for use in working with the Russian Science Citation Index]. Otchjot o nauchno-issledovatel'skoj rabote (promezhutochnyj) po teme «Razrabotka sistemy statisticheskogo analiza rossijskoj nauki na osnove dannyh rossijskogo indeksa citirovanija». — M., 2005 (in Russian).
[7]. Meho L (Meho, Lokman); Yang K (Yang, Kiduk). Fusion approach to citation-based quality assessment. Proceedings Of Issi 2007: 11th International Conference Of The International Society For Scientometrics And Informetrics, Vols I And II : 568-581.
[8]. Angela Vorndran, Alexander Botte. Analysis and evaluation of existing methods and indicators for quality assessment of scientific publications. http://www.eerqi.eu/sites/default/files/Analysis_and_evaluation_of_existing_methods_and_indicators.pdf [PDF].
[9]. EERQI – European Educational Research Quality Indicators. www.eerqi.eu.
[10]. EERQI Project Final Report (2011). http://eerqi.eu/sites/default/files/Final_Report.pdf [PDF].
[11]. Moyses Szklo. Quality of scientific articles. Revista Saúde Pública vol.40 special issue São Paulo Aug. 2006.
[12]. Dr Navneet Gupta BSc (Hons) PhD MCOptom FBCLA. How to Evaluate a Scientific Research Article. http://www.optometry.co.uk/uploads/articles/ARTICLE%200309.pdf [PDF].
[13]. Google Scholar. http://scholar.google.ru.
[14]. Berry M.W., Dumais S.T., O'Brien G.W. Using Linear Algebra for Intelligent Information Retrieval. University of Tennessee Knoxville. TN. USA, 1994.
[15]. Lee D.D., Seung H.S. Learning the parts of objects by non-negative matrix factorization. Nature, 401, pp. 788-791, 1999.
[16]. Rakesh P., Shivapratap G., Divya G., Soman KP. Evaluation of SVD and NMF Methods for Latent Semantic Analysis. International Journal of Recent Trends in Engineering, Vol. 1, No. 3, 2009.
[17]. Griffiths T L, Steyvers M. Finding scientific topics. In: Proceedings of the National Academy of Sciences. USA, 2004, 101: 5228–5235.
[18]. Steinberger J., Ježek K. Text Summarization and Singular Value Decomposition. In Lecture Notes for Computer Science vol. 2457, Springer-Verlag, 2004, pp. 245-254.
[19]. Steinberger J. Text Summarization within the LSA Framework. PhD Thesis, University of West Bohemia in Pilsen, Czech Republic, January 2007.
[20]. Mashechkin I.V., Petrovskij M.I., Carjov D.V. Metody vychislenija relevantnosti fragmentov teksta na osnove tematicheskih modelej v zadache avtomaticheskogo annotirovanija [Methods for calculating the relevance of text fragments on the basis of thematic patterns in the problem of automatic annotation]. Vychislitel'nye metody i programmirovanie. Tom 14, 2013. 91-102 [in Russian].
[21]. Mashechkin I.V., Petrovskiy M.I., Popov D.S., Tsarev D.V. Automatic text summarization using latent semantic analysis. Programming and Computer Software, pp. 299-305, 2011.
[22]. Tsarev D., Petrovskiy M., Mashechkin I. Using NMF-based text summarization to improve supervised and unsupervised classification. 11th International Conference on Hybrid Intelligent Systems (HIS), Malacca, MALAYSIA. P. 185-189, 2011.
[23]. Dmitry Tsarev, Mikhail Petrovskiy and Igor Mashechkin, Supervised and Unsupervised Text Classification via Generic Summarization International Journal of Computer Information Systems and Industrial Management Applications. MIR Labs, Volume 5, 2013, pp. 509-515.

[24]. Wei Xu, Xin Liu, Yihong Gong Document clustering based on non-negative matrix factorization. Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, Toronto, Canada, 2003.
[25]. Y. Ding. Applying weighted PageRank to author citation networks. In Proceedings of JASIST. 2011, pp. 236-245.
[26]. M. Potthast, T. Gollub, M. Hagen, J. Graßegger, J. Kiesel, M. Michel, A. Oberländer, M. Tippmann, A. Barrón-Cedeño, P. Gupta, P. Rosso, B. Stein. Overview of the 4th International Competition on Plagiarism Detection. CLEF2012. 2012.
[27]. S. Alzahrani, N. Salim. Fuzzy Semantic-Based String Similarity for Extrinsic Plagiarism Detection, Lab Report for PAN at CLEF2010, 2010.
[28]. A. Martins. String kernels and similarity measures for information retrieval. 2006.
[29]. Berry M.W., Browne M., Langville A.N., Pauca V.P., Plemmons R.J. Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics and Data Analysis, pp. 155-173, 2007.
[30]. Fomenko V.P., Fomenko T.G. Avtorskij invariant russkih literaturnyh tekstov [Author invariant Russian literary texts]. Predislovie A.T. Fomenko.. Fomenko A.T. Novaja hronologija Grecii: Antichnost' v srednevekov'e. T. 2. M.: Izd-vo MGU, 1996, c.768-820 (in Russian].
[31]. Braslavski P. Document Style Recognition Using Shallow Statistical Analysis. In Proceedings of the ESSLLI 2004 Workshop on Combining Shallow and Deep Processing for NLP, Nancy, France, 2004, p. 1–9.
[32]. DuBay, W.H. The Principles of Readability. Costa Mesa, CA: Impact Information. 2004.
[33]. P.V. Rao and L.L. Kupper, "Ties in paired-comparison experiments: A generalization of the Bradley–Terry model", Amer. Statist. Assoc, 62, 1967, pp. 194–204.
[34]. Turner, H and Firth, D (2012). Bradley-Terry Models in R: The BradleyTerry2 Package. Journal of Statistical Software 48(9), 1–21.
[35]. Hastie, Tibshirani and Friedman (2008). The Elements of Statistical Learning (2nd edition) Springer-Verlag. 763 pages.
[36]. Twisted Framework. http://twistedmatrix.com.
[37]. qooxdoo. http://qooxdoo.org.
[38]. Konferencija «Matematicheskie metody raspoznavanija obrazov» [The Conference «Mathematical Methods of Pattern Recognition»]. http://www.mmro.ru (In Russian).
[39]. The IEEE International Conference on Data Mining (ICDM). http://www.cs.uvm.edu/~icdm.
[40]. International Conference on Mechanical and Electrical Technology (ICMET). http://www.icmet.ac.cn.
[41]. Zhang M.-L., Zhou Z.-H. A k-nearest neighbor based algorithm for multi-label classification. Proceedings of the 1st IEEE International Conference on Granular Computing (GrC'05). Beijing, China, 2005. pp. 718-721.