Прогнозирование цитирования и импакт-фактора терминов для научных публикаций с помощью алгоритмов машинного обучения

А.А. Клоков[1], Е.А. Слободюк[1], М.М. Шарнин[2]
aaklokov@yandex.ru | eslobodyuk@gmail.com | 1@keywen.com
[1]Московский физико-технический институт (НИУ), Москва, Россия,
[2]Институт Проблем Информатики ФИЦ ИУ РАН,

Объектом исследования при написании работы послужил корпус текстовых данных, собранных вместе с научным руководителем и алгоритмы обработки естественного языка анализа. Поток гипотез был проверен на текстах научных публикаций по компьютерным наукам (computer science) с помощью ряда экспериментов по моделированию, описанных в этой диссертации. Предметом исследования являются алгоритмы и результаты работы алгоритмов, направленные на предсказание перспективных тем и терминов, появляющихся в процессе времени в научной среде.

Результатом данной работы является совокупность моделей машинного обучения, с помощью которых проведены эксперименты по выявлению перспективных терминов и семантических связей в корпусе текста.

Полученные модели могут быть использованы для семантической обработки и анализа других предметных областей.

Ключевые слова: *цитирование, импакт-фактор, научные публикации, машинное обучение.*

Predicting the citation and impact factor of terms for scientific publications using machine learning algorithms

A.A Klokov[1], E.A Slobodyuk[1], M.M. Charnine[2]
aaklokov@yandex.ru | eslobodyuk@gmail.com | 1@keywen.com
[1]Moscow Institute of Physics and Technology (NRU), Moscow, Russia
[2]Institute of Informatics Problems RC CSC RAS

The object of the research when writing the work was the body of text data collected together with the scientific advisor and the algorithms for processing the natural language of analysis. The stream of hypotheses has been tested against computer science scientific publications through a series of simulation experiments described in this dissertation. The subject of the research is algorithms and the results of the algorithms, aimed at predicting promising topics and terms that appear in the course of time in the scientific environment.

The result of this work is a set of machine learning models, with the help of which experiments were carried out to identify promising terms and semantic relationships in the text corpus.

The resulting models can be used for semantic processing and analysis of other subject areas.

Key words: *citation, impact factor, scientific publications, machine learning.*

1. Введение

Сегодня при решении важной задачи человек выступает аналитиком, который просматривает огромный объем текстовой информации, например, в сети Интернет. Чтобы принять решение, человек должен проанализировать большое количество ресурсов, потратив достаточное количество времени. Добавим, что каждый год количество текстовой информации растет огромными темпами, в частности научные публикации. Для любой группы исследователей важно быть в курсе свежих научных статей, новых терминов и быстро находить качественную информацию на заданную тему. Возникает идея автоматизировать поиск свежих релевантных (или влиятельных) научных статей на заданную тему. Выявление влиятельных статей уже давно рассматривается как одна из наиболее важных проблем в интеллектуальном анализе данных для научной литературы. Поиск соответствующих научных публикаций является первым и важным этапом для всех научно-исследовательских областей. Однако с быстрым развитием науки и техники ежегодно публикуется огромное количество статей в разных исследовательских областях. Например, в области компьютерных наук в последние годы произошло экспоненциальное увеличение объема публикаций [1]. Все существующие публикации нецелесообразно и невозможно отслеживать, поэтому исследователи должны иметь возможность выделять из информационного поля высококачественные публикации, для этого всю аналитику новых терминов и сравнение текстовых источников можно передать компьютеру, который поможет находить интересные тенденции и идеи в публикациях.

Целью данной работы было освоение теоретических основ и построение моделей машинного обучения, в частности, нейронных сетей, с помощью которых можно исследовать большую выборку научной текстовой информации на одну определенную тему. Например, разделять статьи на классы и подтемы, выделять отдельные кластеры, находить синонимичные фразы между статьями, исследовать влияние некоторых терминов на цитируемость статьи. В нашем случае было решено собрать научные публикации на тему Computer Science за бóльшую часть XX века и опробовать наши модели на этом корпусе текстов. На основе созданных моделей можно видеть развитие тем во времени и другие полезные закономерности. Также создана модель, предсказывающая вероятность того, что определенная статья будет полезна.

2. Основные определения

NLP (англ. Natural Language Processing) - обработка естественного языка с помощью алгоритмов искусственного интеллекта.

Токен - объект, основная единица текста, который может являться словом или последовательностью слов в предложении. Токен отделяется от других токенов знаками препинания или пробелами. Поэтому токенизация – это выделение последовательностей символов, которые могут являться отдельные слова и числа в тексте.

Корпус текстов - некоторый объем текстовой информации, выраженный в нашем случае научными публикациями, которые были заранее подобраны и которые используются для обучения алгоритма-модели.

Обучение алгоритма - настройка его внутренних параметров для выполнения определенной задачи. Настройка весов производится с помощью итеративного изменения их величин в сторону уменьшения ошибок алгоритма по некоторой метрике качества.

Вектор слова - объект в векторном пространстве с некоторой фиксированной размерностью, которому сопоставляется слово из словаря. При определенных обстоятельствах его называют эмбеддингом слова (word embending).

N-грамма - последовательность из n токенов.

Нейро́нная сеть (искусственная нейронная сеть, NN) - это математическая и статистическая модель, представитель большого класса алгоритмов машинного обучения, реализованная на некотором языке программирования, работа которой напоминает принцип работы нервных скоплений клеток живого организма.

Рекуррентная нейронная сеть (RNN) - нейронная сеть, которая специализируется на обработке некоторых последовательностей объектов. В нашем случае – последовательности токенов в предложении.

Семантическая языковая модель (Language Modeling, LM) - нейросетевая модель машинного обучения, которая умеет моделировать связную последовательность токенов (слов, n-грамм). В ней хранятся статистические и семантические зависимости последовательности токенов корпусного словаря. Приведем пример: возьмем всеми известную фразу “Мама мыла раму …”. Обученная языковая модель математически смоделирует, что после входного слова “мама” с большей вероятностью (среди других возможных слов) идет слово “мыла”, а после входных слов “мама мыла” – “раму”, и так далее. То есть, если на вход алгоритму мы подадим вектор слова “мама”, то на выходе получим вектор, близкий к вектору слова “мыла”. Если мы подадим алгоритму вектора слов “мама” и “мыла”, то на выходе получим вектор, близкий к вектору слова “раму”. $V_{n+1}=F(V_0,V_1,\ldots,V_{n-1},V_n)$. Работа LM продемонстрирована на рисунках 1 и 2.

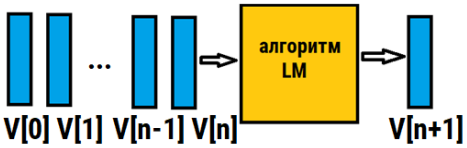

Рис. 1. Работа LM

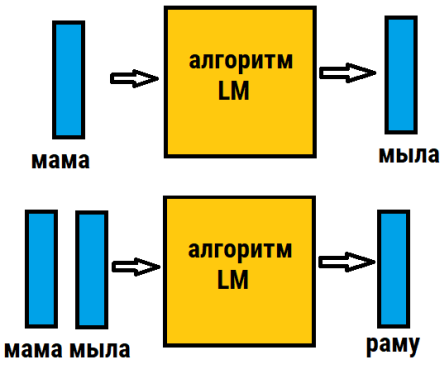

Рис. 2. Работа LM (продолжение)

Нейроны (искусственные нейроны) - базовые единицы нейронной сети прямого распространения, у которых есть входы, внутренние параметры, выход и внутреннее линейное преобразование с дополнительной функцией активации. Строение нейрона продемонстрировано на рисунке 3.

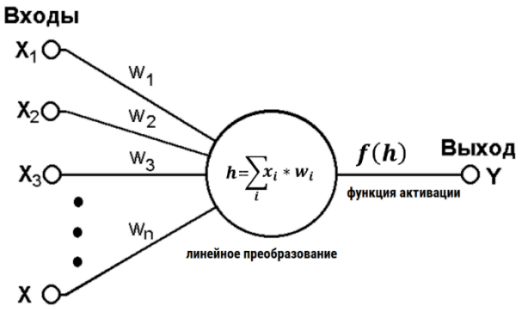

Рис. 3. Строение нейрона

Рекуррентные нейроны - базовые единицы рекуррентной нейронной сети, которые на вход получают не только текущий токен, но и прошлое состояние, которое получилось при отработке предыдущего токена. Строение рекуррентного нейрона продемонстрировано на рисунке 4.

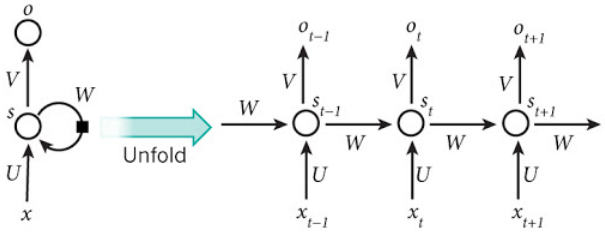

Рис. 4. Строение рекуррентного нейрона

Нормализация – приведение токенов времени, дат, сокращений и так далее к некоторому стандартному виду.

Лемматизация - приведение слов к одинаковой форме: существительные переводятся в единственное число, мужской род, именительный падеж; глаголы, причастия и деепричастия – в инфинитив глагола; прилагательные – в именительный падеж единственного числа в мужском роде.

Стемминг - удаление из слова некоторых изменяемых морфем, например, приставок, окончаний, суффиксов и так далее. Это приводит к нахождению корня/основы слова.

TF-iDF Encoder - статистическая модель, которая переводит текстовое предложение в вектор чисел.

Задача классификации - одна из задач машинного обучения с учителем (supervised learning), при которой по ряду признаков объекта нужно предсказать его класс, к которому он принадлежит.

Задача регрессии - одна из задач машинного обучения с учителем (supervised learning), при которой по ряду признаков объекта нужно предсказать некоторую количественную величину.

Задача кластеризации - одна из задач машинного обучения без учителя (unsupervised learning), при которой по ряду признаков объектов нужно разделить их на некоторое количество совокупностей - кластеров. Объект из одного кластера ближе всего лежит к другим объектам из этого же кластера, чем к объектам из других кластеров.

Импакт-фактор термина - признак слова и метрика, которая показывает количественную меру перспективности термина, который актуален в настоящий момент и который в наибольшей степени влияет на цитирование. ИФТ некоторого термина T на определенный год N - это среднее количество ссылок, приведенных в статьях, опубликованных в течение года N, на статьи с термином T за некоторый период. ИФТ схож с импакт-фактором (ИФ, IF) научных журналов, который приобрел важную роль в оценке результатов работы ученых, кафедр и целых учреждений. Идея применения ИФТ в том, что ИФТ достаточно стационарен на отрезке в один год.

Положительные свойства импакт-фактора:

- благодаря динамическому характеру, ИФТ способен улавливать тенденции и вариации влияния научного результата ученых во времени, в отличие от h-индекса, который является растущей мерой с учетом всей карьеры;
- простота в понимании.

На рисунке 5 представлены графики среднего количества свежих ссылок цитирования на статьи с некоторым термином в зависимости от общего количества (от 2-х до 100) статей с этим термином. Различными цветами отображены графики с различными трендами начальной частоты, т.е. с различным количеством лет от первой статьи с некоторым термином до пятой статьи с этим термином. Ряд1 – 0 лет, Ряд2 – 1 год, Ряд3 – 2 года, ..., Ряд7 – 6 лет, Ряд8 – 7-9 лет, Ряд9 – 10-14лет, Ряд10 – 15-19лет, Ряд11 > 19лет.

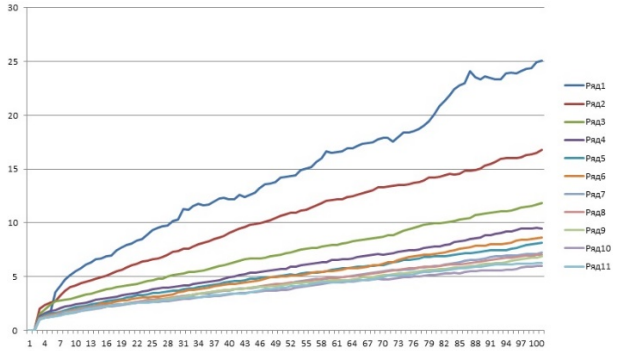

Рис. 5. Графики среднего количества свежих ссылок цитирования

Импакт-фактор научного журнала - метрика и численный показатель цитируемости статей, опубликованных в некотором отдельно взятом журнале. Расчёт импакт-фактора основан на трёхлетнем периоде. Например, импакт-фактор некоторого научного журнала в N-том году есть отношение величин X/Y, где: X - количество ссылок в течение N года на статьи, которые были опубликованы в данном журнале в N-2 и N-1 годах, Y - общая совокупность статей, напечатанных в этом журнале в N-2-N-1 годах (рисунок 6). Чтобы вычислить влияние журналов на цитирование, был введен импакт-фактор (IF) [2]. Эта мера, которая рассчитывается Thomson Reuters [3] каждый год для каждого журнала базы данных Web of Knowledge [4].

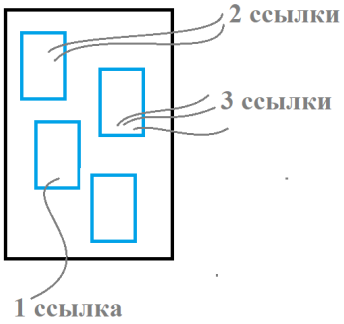

Рис. 6. Импакт-фактор

3. Обзор исследований в обработке естественного языка и оценки качества статей

3.1 Обзор подходов к применению машинного обучения в обработке естественного языка

Сейчас мы находимся на пике интереса к обработке естественного языка (natural language processing, NLP) – области информатики, автоматизирующей работу вычислительной машины с печатным текстом. Благодаря резкому увеличению вычислительных мощностей в течение последних 10 лет стало возможным обрабатывать огромные массивы текста и применять к ним алгоритмы машинного обучения (МО). В обработке естественного языка можно выделить несколько основных направлений:

1. Анализ текста:
   - Извлечение информации.
   - Информационный поиск.
   - Анализ высказываний.
   - Анализ тональности текста.
2. Вопросно-ответные системы.
3. Генерирация текста.
4. Автоматическое выделение основной идеи (summarization) и другие.

Данная работа и описанные ниже эксперименты относятся к анализу текстовой информации. Для этого некоторый объект текста (это может быть слово, предложение или целая научная статья) нужно преобразовать в вектор. Далее компьютер будет работать с векторами текстовых объектов. На сегодняшний день ведущими учеными и исследователями было предложены несколько подходов к векторизации текстовых объектов, которые мы рассмотрим далее.

Один из первых подходов к векторизации текста – мешок слов (BoW – bag of words). Представим, что нам нужно превратить предложение в вектор этим методом. Возьмем большой размерности вектор. Длина вектора – количество слов в словаре (то есть в тексте, в котором мы рассматриваем взятое предложение). Координаты вектора – натуральные числа, соответствующие каждому из слов в этом предложении и показывающий количество таких слов в предложении.

TF-iDF очень похож на мешок слов, но в координатах вектора будет стоять отношение частоты термина в корпусе (отношение числа вхождений некоторого слова к общему числу слов документа) к обратной частоте встречаемости этого слова в других предложениях (частоте употребления слова во всех документах коллекции). Немного подробней, TF-iDF (от англ. TF - term frequency, iDF - inverse document frequency). Таким образом, iDF уменьшает вес широкоупотребительных слов. Для каждого уникального слова в пределах конкретной коллекции документов существует только одно значение iDF. Ввел эту величину Карен Спарк Джонс (рисунок 7).

Рис. 7. Схема TF-iDF

Ранее мы описали методы, которые преобразуют предложение в вектор. Теперь поговорим о методах, преобразующих каждое слово в некоторый вектор, или эмбеддинг. Word2Vec (w2v) – нейросетевой алгоритм, который моделирует векторное пространство слов. На вход ему нужен корпус текстов, а на выходе он выдает набор эмбеддингов слов. То есть, он отображает пространство с высокой размерностью (длина словаря) в пространство небольшой размерности, например 300. Принцип работы этой модели (и многих других моделей семантического анализа) основывается на гипотезе компактности (близкие по смыслу объекты будут лежать ближе друг к другу в некотором пространстве, чем объекты разных семантических тем). Приведем пример: слова «клавиатура» и «монитор» чаще употребляются в одном контексте, нежели «клавиатура» и «помидоры». Математически, модель w2v – нейронная сеть прямого распространения, которая учится прогнозировать слова, которые могут стоять рядом. Здесь можно выделить 2 подхода: прогнозировать слово по его окружению (CBOW - Continuous Bag of Words – рисунок 8) или прогнозировать контекст по центральному слову (Skip-gram – рисунок 9). За величину охвата контекста отвечает “окно” – числовой гиперпараметр, который показывает, сколько слов слева и справа от центрального являются его окружением. Ввел эту модель Томас Миколов в статье 2014 года [5]

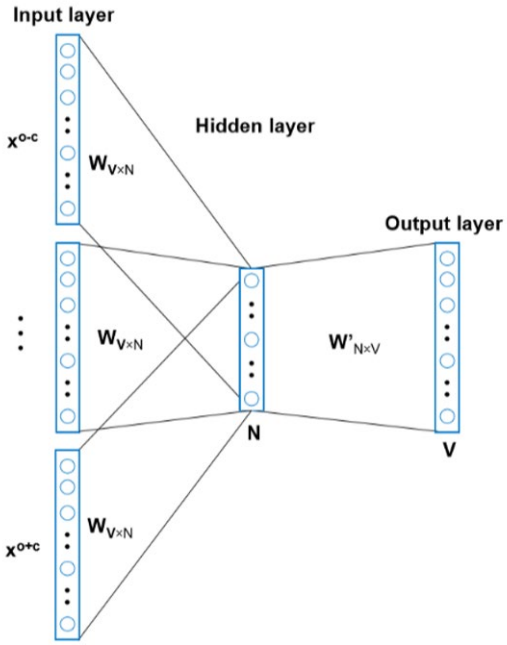

Рис. 8. Схема CBOW

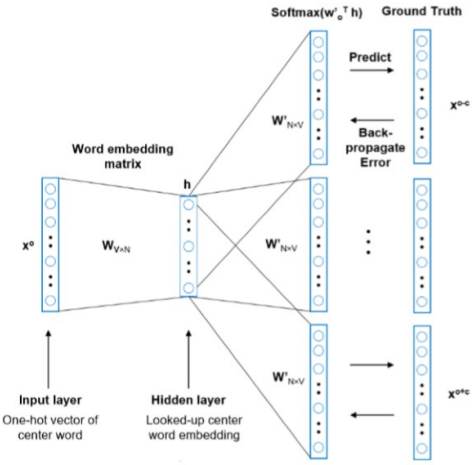

Рис. 9. Схема Skip-gram

Рекуррентные нейронные сети (РНС, англ. Recurrent neural network; RNN) - вид нейронных сетей, где связи между элементами образуют направленную последовательность. Благодаря этому появляется возможность обрабатывать серии событий во времени или последовательные пространственные цепочки. В отличие от многослойных перцептронов, рекуррентные сети могут использовать свою внутреннюю память для обработки последовательностей произвольной, но конечной длины. Поэтому сети RNN применимы в таких задачах, где нечто целостное разбито на части, например слова в тексте. Было предложено много различных архитектурных решений для рекуррентных сетей от простых до сложных. В последнее время наибольшее распространение получили сеть с долговременной и кратковременной памятью (LSTM) и управляемый рекуррентный блок (GRU).

Далее рассмотрим надстройку над рекуррентными нейросетями, которая называется механизмом внимания (в англоязычной литературе - Attention). Механизм внимания был предложен командой Дмитрия Богданова и в статье “Neural Machine Translation by

Jointly Learning to Align and Translate", но использовался для моделей вида энкодер-декодер для решения проблемы ограничения длины входной последовательности одним вектором фиксированной длины. Мы будем использоавать механизм мнимания, так как планируем аналогично обрабатывать последовательности большой длины.

В 2018 году механизм внимания был улучшен, исследователи ввели понятие self-Attention и предложили полностью отказаться от использования рекуррентных слоев в нейронных сетях, а оставить только слои с "улучшенным" вниманием. В статье "Attention is all you need" [6] авторы ввели новую архитектуру нейронной сети - трансформер, разновидности которого применяются и в наши дни. Точность прогнозов немного возросла, а вот объем моделей трансформеров сильно возрос, что накладывает некоторые ограничения их на использование.

3.2. Обзор исследований в области оценки качества научных публикаций и их влияния.

В последние годы ряд исследователей занимался предсказанием будущего цитирования. Их исследования отличались в первую очередь выбором функций, используемых для прогнозирования. Функции можно разделить на те, которые относятся к самой статье, авторам статьи, и другие - журналу, который принял статью для публикации [7]. В настоящее время из-за возможности извлекать данные из онлайновых цифровых библиотек исследования по цитированию проводятся на крупных наборах данных. Модели прогнозирования также стали более сложными из-за успехов машинного обучения. В исследованиях предлагаются самые разные наборы прогностических функций, которые, однако, базируются на метаданных (данные об авторах, месте публикации). Мало исследований, которые включают все содержимое документов. Например, в [8] при моделировании прогнозирования цитирования использовались только первые 200 слов абстракта каждого документа. Многие статьи цитируются двояко - и по номеру в библиографии, и по принципу фамилия–год. При подсчете цитирования не учитывается отношение автора к цитируемым работам (позитивное, негативное), семантика ссылки (сравнение, фактическая информация, определение и т.д.), контекст ссылки (гипотеза, анализ, результат и т.д.) и мотивы цитирования [9], популярность тем, различие типов научных статей (теоретическая статья, экспериментальная статья, позиционная, обзор) [10]. Использование текущих показателей эффективности исследовательских публикаций (Bibliometrics, Altmetrics, Webometrics и т. д.) основано на ложной предпосылке о том, что влияние (или даже качество) исследовательского документа можно оценить исключительно на основе внешних данных без рассмотрения текста самой публикации [11]. Есть некоторые свидетельства того, что в последние годы исследования становятся более междисциплинарными [12], и что это междисциплинарность году, h-индекс в 2005 году менее важен, чем количество документов, опубликованных в 2004-2005 годах.

В [13] при предсказании цитирования использовали систему, которая была основана на вероятностном латентном семантическом анализе (PLSA). Авторы [14] заменили PLSA на LDA. В [15] изучались структурные свойства графов цитирования в информатике. Основное предположение - многие цитаты являются свидетельством информационного потока от одной статьи и ее авторов к другой. Графы цитирования для оценки будущего цитирования использовались в [16]. Альтметрика - альтернатива использованию подсчета цитирования в качестве показателя для оценки воздействия статьи, исследователя или журнала [17] с целью измерить «воздействие» публикации не только в сетях цитирования, но и посредством современных факторов, как Интернет, социальные сети, блоги, комментарии и аннотации к работе. Как указано в [18], ценность публикаций в отношении оценки воздействия может быть оценена посредством полного анализа содержания. Подход авторов опирался на широкий круг экспертов для обзора контента в очень конкретной области, но метод требовал много времени для выполнения.

Авторы [19] проводили полный контент-анализ статей, используя скрытый семантический анализ полных текстов, показав, что среднее расстояние между цитирующими и цитируемыми статьями неуклонно растет с 1990 г. Они использовали LSA для сокращения отношений term-document в корпусе, каждая статья была представлена как вектор LSA в пространстве, расстояние между статьями измерялось путем вычисления косинусной меры. Эта общая тенденция к увеличению расстояния цитирования также совпадает с ростом междисциплинарных исследований [20]. Отмечается, что междисциплинарность вознаграждается несколько более высоким воздействием [21]. В [22] использовалось стандартное косинус-сходство, чтобы рассчитать семантическое сходство между цитирующим документом и цитируемым текстом, затем его оценивали эксперты. Авторы извлекали два предложения до и два предложения после цитаты в оригинальном документе и пытались сопоставить слова в этих предложениях с целевым документом, используя метрику подобия косинуса (рисунок 10).

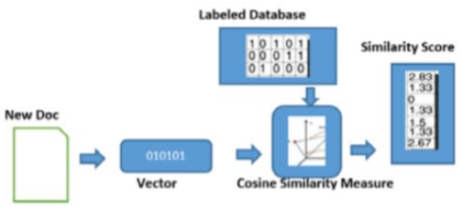

Рис. 10. Метрика подобия косинуса

Для 100 пар с оценкой подобия косинуса более 0,90 система зафиксировала среднюю точность 64,3% на основе оценок экспертов [22]. На рисунке 1 представлена схема расчета оценки сходства (оценка подобия косинуса вычисляет значение, скорректированное для длины статьи, для отображения сходства для каждой

пары предложений на основе значений общих терминов).

Следовательно, существует методологический разрыв в попытках измерить воздействие научной статьи. С одной стороны, широко применяется анализ цитирования, который можно легко отслеживать и автоматизировать, но подобный анализ опирается на многие допущения, которыми легко манипулировать. С другой стороны, экспертная группа проводит более тщательный полный анализ контента, но требуется наличие экспертов, а также огромное количество времени для выполнения.

4. Принцип работы методов, которые были выбраны для проведения экспериментов с собранными данными

4.1 Линейные модели для классификации на примере SVM

Рассмотрим метод опорных векторов (англ. SVM, Support Vector Machine). Будем решать задачу бинарной (когда класса всего два) классификации. Сначала алгоритм тренируется на объектах из обучающей выборки, для которых заранее известны метки классов. Далее уже обученный алгоритм предсказывает метку класса для каждого объекта из отложенной/тестовой выборки. Метки классов могут принимать значения $Y=\{-1,+1\}$. Объект - вектор с N признаками $x=(x_1,x_2,\ldots,x_n)$ в пространстве R^n. При обучении алгоритм должен построить функцию $F(x)=y$, которая принимает в себя аргумент x - объект из пространства R^n и выдает метку класса y.

Задача классификации относится к обучению с учителем. SVM - алгоритм обучения с учителем. Нужно добавить, что SVM может применяться и для задач регрессии, но в данной работе нам понадобится SVM-классификатор.

Главная цель SVM как классификатора - найти уравнение разделяющей гиперплоскости $w_1x_1+w_2x_2+\ldots+w_nx_n+w_0=0$ в пространстве R^n, которая бы разделила два класса неким оптимальным образом. Общий вид преобразования F объекта x в метку класса Y: $F(x)=\text{sign}(w^Tx-b)$. Будем помнить, что мы обозначили $w=(w_1,w_2,\ldots,w_n)$, $b=-w_0$. После настройки весов алгоритма w и b (обучения), все объекты, попадающие по одну сторону от построенной гиперплоскости, будут предсказываться как первый класс, а объекты, попадающие по другую сторону - второй класс.

Внутри функции sign() стоит линейная комбинация признаков объекта с весами алгоритма, именно поэтому SVM относится к линейным алгоритмам. Разделяющую гиперплоскость можно построить разными способами, но в SVM веса w и b настраиваются таким образом, чтобы объекты классов лежали как можно дальше от разделяющей гиперплоскости. Другими словами, алгоритм максимизирует зазор (англ. margin) между гиперплоскостью и объектами классов, которые расположены ближе всего к ней. Такие объекты и называют опорными векторами (см. рис.2). Отсюда и название алгоритма.

Плюсы алгоритма SVM:

1) хорошо работает с пространством признаков большого размера;
2) так алгоритм максимизирует разделяющую полосу, которая, как подушка безопасности, позволяет уменьшить количество ошибок классификации;
3) так как алгоритм сводится к решению задачи квадратичного программирования в выпуклой области, то такая задача всегда имеет единственное решение (разделяющая гиперплоскость с определенными гиперпараметрами алгоритма всегда одна).

Перейдем к выводу правил настройки весов SVM. Чтобы разделяющая гиперплоскость как можно дальше отстояла от точек выборки, ширина полосы должна быть максимальной. Вектор w - вектор нормали к разделяющей гиперплоскости. Здесь и далее будем обозначать скалярное произведение двух векторов a и b - как a^Tb. Давайте найдем проекцию вектора, концами которого будут являться опорные вектора разных классов на вектор w. Эта проекция и будет показывать ширину разделяющий полосы (рисунок 11):

$$\langle (x_+ - x_-), w/\|w\| \rangle = (\langle x_+, w \rangle - \langle x_-, w \rangle)/\|w\| = ((b+1)-(b-1))/\|w\| = 2/\|w\|$$

$$2/\|w\| \to max$$

$$\|w\| \to min$$

$$(w^Tw)/2 \to min$$

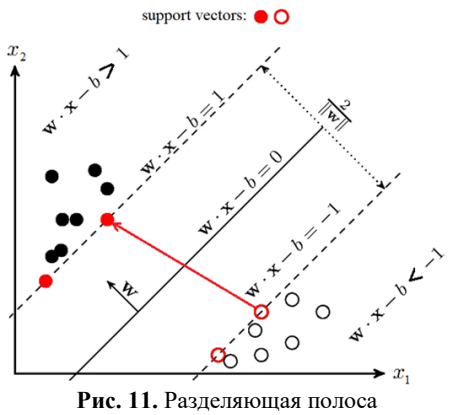

Рис. 11. Разделяющая полоса

Отступом (англ. margin) объекта x от границы классов называется величина $M=y(wTx-b)$. Алгоритм допускает ошибку на объекте тогда и только тогда, когда отступ M отрицателен (когда y и $(wTx-b)$ разных знаков). Если M принадлежит интервалу (0;1), то объект попадает внутрь разделяющей полосы. Если M>1, то объект x классифицируется правильно, и находится на некотором удалении от разделяющей полосы. Т.е. алгоритм будет правильно классифицировать объекты, если выполняется условие:

$$y(w^Tx - b) \geqslant 1$$

Если объединить два выведенных выражения, то получим базовую настройку SVM с жестким зазором (hard-margin SVM), когда никакому объекту не разрешается попадать на полосу разделения. Решается

аналитически через теорему Куна-Таккера. Получаемая задача эквивалентна двойственной задаче поиска седловой точки функции Лагранжа.

$$\begin{cases} (w^T w)/2 \to min \\ y(w^T x - b) \geqslant 1 \end{cases}$$

Из этой системы можем вывести функцию оптимизации, Q - функция потерь, она же loss function.

$$Q = max(0, 1 - yw^T x) + \alpha(w^T w)/2$$

Именно ее мы и будем минимизировать с помощью градиентного спуска. Выведем правила изменения весов, где η – шаг спуска:

$$w = w - \eta \bigtriangledown Q$$

$$\bigtriangledown Q = \begin{cases} \alpha w - yx & \text{, если } yw^T x < 1 \\ \alpha w & \text{, если } yw^T x \geqslant 1 \end{cases}$$

Изменяя веса w в сторону уменьшения функции Q, мы обучаем алгоритм делать правильные прогнозы, совершая всё меньше ошибок.

4.2 Модель word2vec для создания эмбеддингов слов

Понятно, что мы не можем подавать алгоритмам на вход слова, состоящие из букв. Нам нужно каждое слово преобразовать в некоторый вектор, то есть получить эмбеддинг этого слова. Word2Vec (w2v) – алгоритм, который моделирует векторное пространство слов. На вход ему нужен корпус текстов, а на выходе он выдает набор эмбеддингов слов. То есть, он отображает пространство с высокой размерностью (длина словаря) в пространство небольшой размерности, например 300. Принцип работы этой модели (и многих других моделей семантического анализа) основывается на гипотезе компактности (близкие по смыслу объекты будут лежать ближе друг к другу в некотором пространстве, чем объекты разных семантических тем). Приведем пример: слова «клавиатура» и «монитор» чаще употребляются в одном контексте, нежели «клавиатура» и «помидоры». Математически, модель w2v – нейронная сеть прямого распространения, которая учится прогнозировать слова, которые могут стоять рядом. Здесь можно выделить 2 подхода: прогнозировать слово по его окружению (CBOW - Continuous Bag of Words) или прогнозировать контекст по центральному слову (Skip-gram). За величину охвата контекста отвечает “окно” – числовой гиперпараметр, который показывает, сколько слов слева и справа от центрального являются его окружением. На приведенных ниже рисунках 12 и 13 окно равно 1.

Рис. 12. CBOW при окне, равном 1

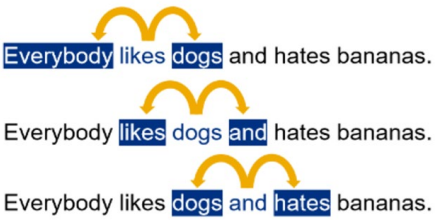

Рис. 13. Skip-gram при окне, равном 1

4.3 Нейросетевые модели для классификации на примере рекуррентной нейронной сети с Attention

Нейронные сети представляют собой несколько слоев линейных преобразований, разделенных нелинейными функциями активации. Обучаются нейронные сети методом обратного распространения ошибки. Для обработки последовательностей отлично подходят рекуррентные слои, в которые можно подать последовательность эмбеддингов слов предложения. После преобразований эмбеддингов слов в рекуррентных слоях получаются некоторые вектора, описывающие признаки входного предложения. Чтобы помочь нейросети концентрироваться на определенных признаках предложения, добавим слой внимания (attention), который нужен, чтобы все внутренние вектора-признаки, вышедшие из рекуррентных слоев, объединились в один финальный вектор. Далее финальный, или результирующий, вектор предложения проходит через несколько слоев прямого распространения для решения задачи классификации.

Рассмотрим более подробно, как работает механизм внимания: чтобы получить финальный вектор предложения, в котором содержится вся информация о предложении, нам нужно сложить все внутренние вектора-признаки с некоторыми коэффициентами a_i (составить линейную комбинацию). Коэффициенты – это числа, которые показывают, насколько похож текущий вектор h_i на последний вектор h_T. Более того, сумма всех коэффициентов равна 1: $sum(a_i) = 1$.

Функция a_i = F(h_i,h_T) может быть разной: это может быть простое косинусное расстояние между двумя векторами, может быть поэлементным произведением двух векторов, или еще одним слоем нейронной сети прямого распространения.

Чтобы сумма всех коэффициентов была равна 1, нам нужно применить скалирующую функцию softmax. Она является нелинейным преобразованием массива полученных коэффициентов:

$$a\_i = \frac{e^{a\_i}}{\sum_i e^{a\_i}}$$

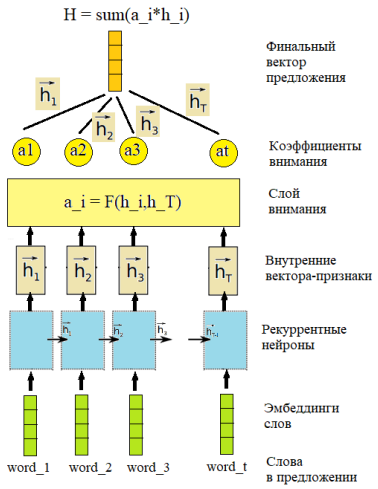

Рис. 14. Схема работы механизма внимания

4.4. Модель линейной регрессии

Теперь рассмотрим метод линейной регрессии. Необходимо задать модель зависимости объясняемой переменной y от объясняющих ее факторов, функция зависимости будет линейной: y=w0+∑i=1mwixi. Сначала алгоритм тренируется на объектах из обучающей выборки, для которых заранее известны правильные ответы. Далее уже обученный алгоритм предсказывает ответ для каждого объекта из отложенной/тестовой выборки. Объект - вектор с N признаками x=(x1,x2,…,xn) в пространстве Rn.

$$\vec{y} = X\vec{w} + \epsilon,$$

где

- $\vec{y} \in \mathbb{R}^n$ – объясняемая (или целевая) переменная;
- w – вектор параметров модели (в машинном обучении эти параметры часто называют весами);
- X – матрица наблюдений и признаков размерности n строк на $m+1$ столбцов (включая фиктивную единичную колонку слева) с полным рангом по столбцам: $\text{rank}\,(X) = m+1$;
- ϵ – случайная переменная, соответствующая случайной, непрогнозируемой ошибке модели.

Задача регрессии, как и задача классификации, относится к обучению с учителем. Главная цель линейной регрессии - провести линию в некотором пространстве признаков рядом/через объекты из тренировочной выборки. При обучении алгоритм должен построить функцию F(x)=y, которая принимает в себя аргумент x - объект из пространства R^n и выдает y. Более строго, общий вид преобразования F объекта x в прогнозируемую величину следующий:

$$y= F(x)=w_1x_1+w_2x_2+\ldots+w_nx_n+w_0+e.$$

5. Проведение экспериментов и результаты

5.1. Предварительная обработка текстов

Перед построением моделей машинного обучения корпус текстов нужно подготовить. Предобработка текста нужна для увеличения скорости обработки текстов и уменьшения затрачиваемой памяти на тексты и словарь. Уменьшение длины словаря приводит к уменьшению общего количества внутренних параметров моделей (то есть они будут быстрее обучаться).

Предобработка текста состоит из нескольких последовательных этапов:

1. Приведение всех букв к нижнему регистру.
2. Исключение стоп-слов – удаление наиболее встречающихся слов в тексте, таких как местоимения, предлоги, союзы, числительные и т.д. Данные слова часто встречаются на протяжении любых текстов и не несут особого смысла, но при этом сильно портят реальную семантическую картину текста т.к. участвуют во многих связях с разными словами и тем самым портят конечный вид матрицы частоты встречаемости отдельных слов. Важно учесть, что в стоп-слова не входят токены “нет”, “не”, “no”, “not”, так как они меняют семантическое значение следующего токена на противоположное.
3. Исключение редко встречающихся слов – удаление слов, которые редко встречались в тексте. Данная процедура необходима для ускорения работы методов, т.к. уменьшает кол-во слов, подлежащих обработке без потери общности конечной семантической картины.
4. Нормализация текста (смотреть пункт 2.1).
5. Лемматизация текста (иногда вместо лемматизации применяется стемминг – смотреть пункт 2.1).
6. Дополнительно можно соединить токены “нет”, “не”, “no”, “not” и подобные со следующим токеном, чтобы образовался один токен вместо двух.

```
dog, dogs, dog's, dogs' => dog
```

```
the boy's dogs are different sizes => the boy dog be differ size
```

5.2. Модель классификации с помощью SVM

В собранном корпусе текстов научных публикаций по теме Computer Science находятся тексты статей, авторы статей, каждая статья может ссылаться на другие. Поэтому на основе корпуса был создан датасет для экспериментов по прогнозу цитируемости статьи. Решалась задача классификации: будут ли на некоторую взятую статью ссылаться какие-либо другие авторы в течение следующих трёх лет (класс 1) или нет (класс 0).

Эксперимент состоял в том, чтобы проверить, насколько хорошо алгоритм машинного обучения SVM сможет прогнозировать цитируемость статьи (отличать класс 1 от класса 0).

Алгоритм обучался на метках классов и соответствующих им названиях публикаций. Чтобы получить признаки из текста – названия статей сначала проходили все стадии предобработки (пункт 3.1), а после предобработки – преобразовывались в вектора с помощью TF-iDF векторайзера. После обучения алгоритм проверялся на отложенной выборке, состоящей из 10% датасета. Метрика, по которой проверялся алгоритм – ROC AUC, составила 0.62. Это неплохой результат, который мы получили, используя только названия статей. После добавления автора – метрика возросла до 0.65.

5.3. Модель классификации с помощью рекуррентной нейронной сети с Attention

Подготовка датасета была аналогичной, как в пункте 3.2. В собранном корпусе текстов научных публикаций по теме Computer Science находятся тексты статей, авторы статей, каждая статья может ссылаться на другие. Поэтому на основе корпуса был создан датасет для экспериментов по прогнозу цитируемости статьи. Решалась задача классификации: будут ли на некоторую взятую статью ссылаться какие-либо другие авторы в течение следующих трёх лет (класс 1) или нет (класс 0).

Эксперимент состоял в том, чтобы проверить, насколько хорошо глубокая рекуррентная нейронная сеть с механизмом внимания сможет прогнозировать цитируемость статьи (отличать класс 1 от класса 0).

Так как эксперименты делались с целью подобрать наилучшую архитектуру, решаю поставленную задачу, то приведем некоторую общую архитектуру (рисунок 2). Изменяемые параметры: n - количество рекуррентных слоев – от одного до четырех (с разным количество нейронов), m - количество слоев прямого распространения – от 1 до 3х с разным количеством нейронов).

Нейросеть (рисунок 15) обучалась на метках классов и соответствующих им названиях публикаций. Чтобы получить признаки из текста – названия статей сначала проходили все стадии предобработки (пункт 3.1), а после предобработки – преобразовывались в вектора с помощью другой нейронной модели - w2v модели. После обучения алгоритм проверялся на отложенной выборке, состоящей из 10% датасета. Получены результаты: для лучше архитектуры нейросети – 0.67 по метрике ROC AUC.

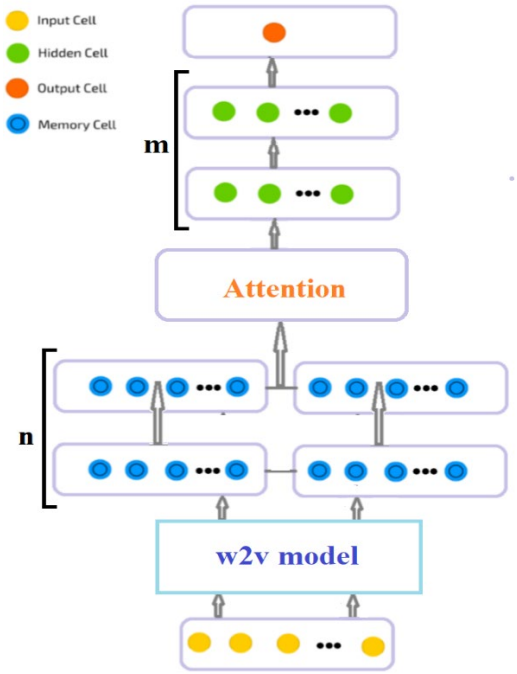

Рис 15. Нейросеть

5.4. Модель линейной регрессии для прогнозирования импакт-фактора терминов

Для экспериментов с введенным выше импакт-фактором терминов (ИФТ) был подготовлен датасет: выбраны определенные слова и выражения, которые чаще всего встречались в названиях сильно цитируемых статей. Для каждого термина были посчитаны величины – признаки:

1. ИФТ на следующий год – величина, которую хотим научиться прогнозировать.
2. Количество статей с термином за текущий год.
3. Тренд - рост количества статей за 2 года (посчитанный для текущего года).
4. ИФТм от внутренних ссылок (посчитанный для текущего года).
5. ИФТ (посчитанный для текущего года).
6. Тренд - рост количества статей за 2 года (посчитанный для предыдущего года).
7. ИФТм от внутренних ссылок (посчитанный для предыдущего года).
8. ИФТ (посчитанный для предыдущего года).
9. Тренд - рост колва статей за 2 года (посчитанный для пред-предыдущего года).
10. ИФТм от внутренних ссылок (посчитанный для пред-предыдущего года).
11. ИФТ (посчитанный для пред-предыдущего года).

Были проведены следующие эксперименты: прогнозирование ИФТ для следующего года по 3 признакам: ИФТ для пред-предыдущего года, ИФТ для предыдущего года и ИФТ для текущего года. По метрике MSE получили результат 0.045, прогнозирование ИФТ для следующего года по всем признакам, описанным выше – по метрике MSE получили результат 0.03.

Приведем графики важности признаков для линейных регрессий (рисунок 16).

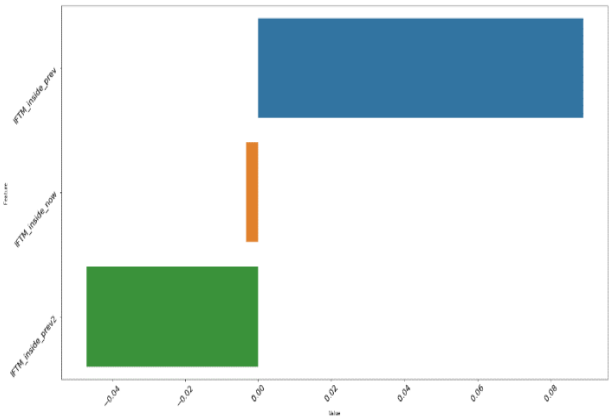

Рис. 16. Графики важности признаков для линейных регрессий

6. Заключение, выводы

В течение работы разработана концепция импакт-фактор терминов, позволяющая более точно прогнозировать библиографические ссылки на статьи. Понимание этого метода приведет к лучшему поиску полезных публикаций. ИФ термина нужен для выявления перспективных тем и идей, которые актуальны в настоящий момент и которые в наибольшей степени

влияют на цитирование. В данной работе протестирован метод предсказания ИФ термина на основе статистических выборок подобных трендовых (популярных) терминов.

Существующие методы оценки влияния научных публикаций либо базируются на рассмотрении метаданных (данных об авторах и месте публикации), либо, при рассмотрении полнотекстовых документов, требуют большой ручной работы экспертов. В отличие от существующих подходов в данной работе из названий публикаций извлекаются информативные токены, которые используются для прогнозирования цитирования и прогнозирование импакт-фактора этих токенов. Отслеживание импакт-фактор информативных токенов приведет к пониманию развития тем, предметных траекторий (траекторий идей) в научном семантическом пространстве.

Можно сделать следующие выводы:

1. Проведенные исследования подтверждают целесообразность применения алгоритмов машинного обучения и технологии нейронных сетей для работы с естественными языками.
2. Приведенные алгоритмы выявляют полезные статьи и термины в их называниях, что позволяет находить решать задачи, связанные с семантическим анализом предметной области и полезен при построении информационного поиска.

Также были выполнены следующие задачи:

- проведен обзор литературы, в которой описываются существующие методы и алгоритмы обработки естественных языков;
- выбраны наиболее подходящие методы и алгоритмы;
- алгоритмы реализованы с целью использования;
- проведены эксперименты по прогнозированию цитирования статей;
- статистически выделены термины и построена модель прогноза их импакт-факторов.

7. Применение

Проблема прогнозирования количества цитирований в научной статье имеет много приложений в разных областях. Исследователи смогут заранее распознать наиболее влиятельные статьи, чтобы делать быстрый информационный поиск и улучшить планирование своих исследований. Идея прогнозирования цитирования и нахождения перспективных терминов области поможет научным группам вести более качественный интеллектуальный информационный поиск, а также приведет к развитию моделей предсказания трендов в любой научной области.

Более того, исследования трендов терминов позволяет использовать импакт-фактор не только в научных статьях, но и для текстов из социальных сетей (где есть похожие механики цитирований или репостов) для отслеживания настроения социальных масс.

Благодарность

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта 19-07-00857.

Список информационных источников

[1] Bethard S., Jurafsky D. Who should I cite: learning literature search models from citation behavior. In Proceedings of the 19th ACM international conference on Information and knowledge management, pages 609-618. ACM, 2010.

[2] Pan, R. K. & Fortunato, S. (2014). Author Impact Factor: tracking the dynamics of individual scientific impact. Scientific reports, 4, 4880. (Импакт-фактор автора: отслеживание динамики индивидуального научного воздействия).

[3] Petersen, A. M. et al. Reputation and Impact in Academic Careers arXiv:1303.7274.

[4] Web of knowledge. http://wokinfo.com/ (2014). Accessed: 2014-02-01.

[5] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality (2013) https://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compo.

[6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention Is All You Need https://arxiv.org/abs/1706.03762.

[7] Yan R., Huang C., Tang J., Zhang Y., Li X. To Better Stand on the Shoulder of Giants, JCDL. 2012. Available at: http://keg.cs.tsinghua.edu.cn/jietang/publications/JCDL12-Yan-et-al-To-Better-Stand-on-the-Shoulderof-Giants.pdf.

[8] Ho Qirong, Eisenstein J., Xing E. P. 2012. Document hierarchies from text and links. In Proceedings of the 21st international conference on World Wide Web, pages 739–748.

[9] Nicolaisen J. Citation Analysis // Annual Review of Information Science and Technology. - 2007. - Vol. 41, No.1. - P. 609-641. - http://doi.org/10.1002/aris.2007. 1440410120.

[10] Seglen P. O. Why the impact factor of journals should not be used for evaluating research// BMJ: British Medical Journal. - 1997.- No. 314 (February).- P. 498-502. - http://dx.doi.org/10.1136/bmj.314.7079.497.

[11] Кнот, П., Германнова, Д. На пути к семантометриям: новый семантический критерий сходства для оценки вклада научной публикации. Междунар. форум по информ. 2015. Т. 40. № 1, с. 3-8.

[12] Porter A. L., Rafols I. 2009. Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics 81, 3: 719–745. http://doi.org/10.1007/s11192-008-2197-2.

[13] Hofmann D., Cohn T. 2001. The missing link a probabilistic model of document content and hypertext connectivity. In Proceedings of the 2000 Conference on Advances in Neural Information

Processing Systems. The MIT Press, pages 430–436.

[14] Erosheva E., Fienberg S., Lafferty J. 2004. Mixedmembership models of scientific publications. Proceedings of the National Academy of Sciences of the United States of America, 101(Suppl.1):5220–5227.

[15] Shi X., Tseng B., Adamic L. Information Diffusion in Computer Science Citation Networks. arXiv:0905.2636v1 [cs.DL] 15 May 2009.

[16] PobiedinaN.,IchiseR.PredictingCitationCountsfor Academic Literature Using Graph Pattern Mining. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent SystemsIEA/AIE2014:ModernAdvancesinApplied Intelligence,.

[17] Priem J., Taraborelli D., Groth P., Neylon C. Altmetrics: A manifesto, 2010. http://altmetrics.org/manifesto/.

[18] Sutherland W. J., Goulson D., Potts S. G., Dicks L. V. Quantifying the impact and relevance of scientific research. PLoS ONE, 6(11), 2011. http://doi.org/10.1371/journal.pone.0027537.

[19] Whalen R., Huang Y. et al. Citation distance: Measuring changes in scientific search strategies. In Proceedings of the 25th International Conference Companion on World Wide Web, WWW '16 Companion, pages 419-423, Republic and Canton of Geneva, Switzer.

[20] Porter A. L., Rafols I. 2009. Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. Scientometrics 81, 3: 719–745. http://doi.org/10.1007/s11192-008-2197-2.

[21] Shiji Chen, Cl´ement Arsenault, Vincent Larivi`ere. 2015. Are top-cited papers more interdisciplinary? Journal of Informetrics 9, 4: 1034–1046. http://doi.org/10.1016/j.joi.2015.09.003.

[22] Alam H., Kumar A., Werner T., Vyas M. Are Cited References Meaningful? Measuring Semantic Relatedness in Citation Analysis. In: Proc. of the 2nd Joint Workshop on Bibliometricenhanced Information Retrieval and Natural Language Processing for Digital.