Тематические модели: учет сходства между униграммами и биграммами

© М. А. Нокель
МГУ им. М. В. Ломоносова, Москва
mnokel@gmail.com

Аннотация

В статье представлены результаты экспериментов по добавлению сходства между униграммами и биграммами в тематические модели. Вначале изучается возможность применения ассоциативных мер для выбора и последующего включения биграмм в тематические модели. Затем предлагается модификация оригинального алгоритма PLSA, учитывающая похожие униграммы и биграммы, начинающиеся с одних и тех же букв. И в конце статьи предлагается новый итеративный алгоритм без учителя, показывающий, как темы сами могут выбирать себе наиболее подходящие биграммы. В качестве текстовой коллекции была взята подборка статей из электронных банковских журналов на русском языке. Эксперименты показывают значительное улучшение качества тематических моделей по всем целевым метрикам.

1 Введение

Вероятностные тематические модели (далее просто *тематические модели*) – одно из современных приложений машинного обучения к анализу текстов. Тематические модели предназначены для описания текстов с точки зрения их тем. Они определяют, к каким темам относится каждый документ в текстовой коллекции и какие слова образуют каждую такую тему. При этом темы представляются в виде дискретных распределений на множестве слов, а документы – в виде дискретных распределений на множестве тем [1]. Пользователям темы предоставляются, как правило, в виде некоторых списков часто встречающихся рядом друг с другом слов, упорядоченных по убыванию степени принадлежности им.

Труды 16-й Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" – RCDL-2014, Дубна, Россия, 13-16 октября 2014 г.

С момента своего появления тематические модели достигли значительных успехов в задачах информационного поиска [2], разрешении морфологической неоднозначности [3], многодокументного аннотирования [4], кластеризации и категоризации документов [5]. Также они успешно применяются в выявлении трендов в научных публикациях и новостных потоках [6], обработке аудио- и видео-сигналов [7] и других задачах. Самыми известными представителями являются латентное размещение Дирихле (LDA) [1], использующее априорное распределение Дирихле, и метод вероятностного латентного семантического анализа (PLSA) [8], не связанный ни с какими параметрическими априорными распределениями.

В работах [9] и [10] было показано, что использование тематических моделей в задаче извлечения однословных терминов способно значительно улучшить качество извлечения последних из текстов предметных областей. Поэтому актуальной является и проблема улучшения качества самих тематических моделей за счет использования некоторой лингвистической информации, чему и посвящена данная работа.

Одним из главных недостатков тематических моделей является использование модели "мешка слов", в которой каждый документ рассматривается как набор встречающихся в нем слов. Данная модель не учитывает порядок слов и основывается на гипотезе независимости появлений слов в документах друг от друга. На данный момент проведено множество исследований, посвященных изучению вопроса добавления словосочетаний, n-грамм и многословных терминов в тематические модели. Однако часто это приводит к ухудшению качества модели в связи с увеличением размера словаря или к значительному усложнению модели [12], [13], [14].

В статье предлагается новый подход, позволяющий учесть взаимосвязь между похожими словами (в частности, однокоренными) в тематических моделях (такими, как *банк – банковский – банкир*, *кредит – кредитный – кредитовать – кредитование*). На основании данного метода в статье описывается и новый подход к добавлению биграмм в тематические модели, который рассматривает биграммы уже не как "черные ящики", а

учитывает взаимосвязь между ними и униграммами, основанную на их внутренней структуре. Предлагаемый алгоритм улучшает качество тематических моделей по двум целевым метрикам: перплексии и согласованности тем [15].

Все эксперименты, описанные в статье, проведены на основе алгоритма PLSA и его модификаций на коллекции текстов банковской тематики на русском языке, взятых из электронных журналов.

Статья организована следующим образом. В разделе 2 рассматриваются близкие работы. В разделе 3 описывается текстовая коллекция, использующаяся в экспериментах, все стадии её предобработки и метрики, применяемые для оценивания качества работы тематических моделей. В разделе 4 проводится обширный анализ ассоциативных мер для выбора и последующего включения биграмм в тематические модели. В разделе 5 предлагается новый алгоритм, позволяющий учесть сходство между униграммами и биграммами в тематических моделях. В разделе 6 предлагается еще один новый итеративный алгоритм, использующий тот факт, что темы могут сами выбирать себе наиболее подходящие биграммы. И в последнем разделе приводятся выводы.

2 Близкие работы

2.1 Тематические модели

На сегодняшний день разработано достаточно много различных тематических моделей. Исторически одними из первых появились модели, основанные на традиционных методах кластеризации текстов [11]. При этом после окончания работы алгоритма кластеризации каждый получившийся кластер рассматривается как отдельная тема для вычисления вероятностей входящих в него слов по следующей формуле:

$$P(w|t) = \frac{f(w|t)}{\sum\limits_{w} f(w|t)}$$

где $f(w|t)$ – частотность слова w в теме t.

Естественным ограничением таких моделей является отнесение каждого документа лишь к одной теме.

В последнее время появились вероятностные механизмы нахождения тем в документах, рассматривающие каждый документ в виде смеси тем, а каждую тему в виде некоторого вероятностного распределения над словами. Вероятностные модели порождают слова по следующему правилу:

$$P(w|d) = \sum_{t} P(w|t)P(t|d)$$

где $P(t|d)$ и $P(w|t)$ – распределение тем по документам и слов по темам, а $P(w|d)$ – наблюдаемое распределение слов по документам.

Согласно данной модели коллекция D – это выборка наблюдений (d, w), генерируемых Алгоритмом 1.

Algorithm 1: Порождение коллекции текстов с помощью тематической модели

Input: распределения $P(w|t)$ и $P(t|d)$
Output: коллекция $D = \{(d, w)\}$

```
for d ∈ D do
    Задать длину n_d документа d
    for i = 1, ..., n_d do
        Выбрать тему t из P(t|d)
        Выбрать слово w из P(w|t)
        Добавить в D пару (d, w)
```

Самыми известными представителями данной категории являются метод вероятностного латентного семантического анализа (PLSA) [8] и латентное размещение Дирихле (LDA) [1].

2.2 Словосочетания в тематических моделях

Все описанные в прошлом разделе алгоритмы работают только со словами, основываясь на гипотезе о независимости слов друг от друга – модели “мешка слов”. Идея же использования словосочетаний в тематических моделях сама по себе не нова. На данный момент существуют 2 подхода к решению данной проблемы: создание унифицированной вероятностной модели и предварительное извлечение словосочетаний и n-грамм для их последующего добавления в тематические модели.

Большинство исследований на данный момент посвящено первому подходу. Так, первая попытка выйти за пределы модели “мешка слов” была предпринята в работе [12], где была представлена Биграммная Тематическая Модель. В этой модели вероятности слов зависят от вероятностей непосредственно предшествующих им слов. Модель словосочетаний LDA расширяет Биграммную Тематическую Модель за счет введения дополнительных переменных, способных генерировать и униграммы, и биграммы. В работе [14] представлена Тематическая N-граммная Модель, усложняющая предыдущие для обеспечения возможности формирования биграмм в зависимости от контекста. В работе [16] предложена тематическая модель Слово-Символ, выходящая за рамки использовавшегося ранее предположения о том, что тема каждой n-граммы определяется в зависимости от тем слов, составляющих данное словосочетание. Эта модель оказалась наиболее пригодной для китайского языка. В работе [17] устанавливается связь между LDA и вероятностными контекстно-свободными грамматиками и предлагаются две но-

вые вероятностные модели, сочетающие в себе идеи из LDA и вероятностных контекстно-свободных грамматик для добавления словосочетаний и имен собственных в тематические модели.

Несмотря на то, что все описанные выше модели имеют теоретически элегантное обоснование, у них очень большая вычислительная сложность, что ведёт к неприменимости на реальных данных. Так, например, вычислительная сложность Биграммной Тематической Модели равна $O(W^2T)$, в то время как для LDA она равна $O(WT)$, для PLSA – $O(WT+DT)$, где W – размер словаря, D – количество документов в коллекции и T – число тем. Поэтому такие модели представляют в основном чисто теоретический интерес.

Алгоритм, предложенный в работе [18], относится ко второму типу методов, добавляющих словосочетания в тематические модели. На этапе предобработки авторы извлекают биграммы с помощью t-теста и заменяют отдельные униграммы лучшими по данной мере биграммами. При этом используются 2 метрики оценивания качества полученных тем: перплексия и согласованность тем [15]. В статье показано, что добавление биграмм в тематические модели приводит к ухудшению перплексии и к улучшению согласованности тем.

Данная работа также относится ко второму типу методов и отличается от работы [18] в том, что описываемый здесь подход учитывает внутреннюю структуру биграмм и взаимосвязь между ними и составляющими их униграммами, что приводит к улучшению обоих показателей: и перплексии, и согласованности тем.

Идея использования априорных лингвистических знаний в тематических моделях сама по себе не нова. Так, в работе [19] предметно-ориентированные знания представляются в виде Must-Link и Cannot-Link примитивов с помощью априорного леса Дирихле. Эти примитивы отвечают за то, чтобы слова порождались одними и теми же или, наоборот, разными темами. Однако позднее было замечено, что данный метод может привести к экспоненциальному росту при кодировании Cannot-Link примитивов, и потому его сложно применять с большим количеством ограничений [20]. Другой способ включения подобных знаний представлен в работе [21], где был предложен частично обучаемый с учителем EM-алгоритм для группировки выражений в некоторые заданные пользователем категории. Для обеспечения наилучшей инициализации EM-алгоритма предложенный в статье метод использует априорное знание о том, что синонимы и выражения, имеющие одинаковые слова, должны, скорее всего, относиться к одним и тем же группам. Данная работа отличается от приведённых выше тем, что в ней сходства между униграммами и биграммами добавляются в тематическую модель естественным образом путем подсчета их совместной встречаемости в документах коллекции. Предлагаемый подход никак не увеличивает вычислительную сложность оригинального алгоритма PLSA.

3 Текстовая коллекция и методы оценивания качества тематических моделей

3.1 Текстовая коллекция и предобработка

В экспериментах, описанных в данной статье, использовалась текстовая коллекция из 10422 статей на русском языке, взятых из некоторых электронных банковских журналов (таких, как Аудитор, РБК, Банковский журнал и др.). В данных документах содержится почти 15.5 млн слов.

На этапе предобработки был проведен морфологический анализ документов. В экспериментах рассматривались только *существительные*, *прилагательные*, *глаголы* и *наречия*, поскольку служебные слова не играют значительной роли в определении тем. Кроме того, из рассмотрения исключались слова, встретившиеся менее 5 раз во всей текстовой коллекции.

На этапе предобработки из документов также извлекались биграммы в формах *сущ. + сущ. в родительном падеже* и *прил. + сущ.* В экспериментах рассматривались только такие биграммы, поскольку темы, как правило, задаются именными группами.

3.2 Методы оценивания качества тематических моделей

Для оценивания качества полученных тем в статье рассматриваются две метрики.

Во-первых, использовалась *перплексия*, являющаяся стандартным критерием качества тематических моделей [22]. Эта мера несоответствия модели $p(w|d)$ словам w, наблюдаемым в документах коллекции, определяется через логарифм правдоподобия:

$$Perplexity(D) = \exp\left(-\frac{1}{n}\sum_{d\in D}\sum_{w\in d} n_{dw}\ln p(w|d)\right)$$

где n – число всех рассматриваемых слов в текстовой коллекции, D – множество всех документов в коллекции, n_{dw} – частота слова w в документе d, $p(w|d)$ – вероятность появления слова w в документе d.

Чем меньше значение перплексии, тем лучше модель предсказывает появление слов w в документах коллекции D. Поскольку известно, что перплексия, вычисленная на той же самой обучающей коллекции документов, склонна к переобучению и может давать оптимистически заниженные

значения [1], в данной статье используется стандартный метод вычисления контрольной перплексии, описанный в работе [24]. Коллекция документов изначально разбивалась на 2 части: обучающую D, по которой строилась модель, и контрольную D', по которой вычислялась данная метрика. Хотя на данный момент существует множество исследований, утверждающих, что перплексию нельзя применять для оценивания качества тематических моделей [23], данная метрика попрежнему широко используется для сравнения различных тематических моделей.

В то же время неоднократно предпринимались попытки предложить способ автоматического оценивания качества тематических моделей, никак не связанного с перплексией и коррелирующего с мнениями экспертов. Данная постановка задачи является очень сложной, поскольку эксперты могут достаточно сильно расходиться во мнениях. Однако в недавних работах [15], [25] было показано, что возможно автоматически оценивать *согласованность тем*, основываясь на семантике слов с точностью, почти совпадающей с экспертами. Предложенная метрика измеряет интерпретируемость тем, основываясь на способах оценивания экспертом [15]. Поскольку темы, как правило, предоставляются экспертам для проверки в виде первых топ-N слов, согласованность тем оценивает то, насколько данные слова соответствуют рассматриваемой теме. Newman в работе [15] предложил использовать автоматический способ вычисления данной метрики исходя из меры взаимной информации:

$$TC\text{-}PMI(t) = \sum_{j=2}^{10} \sum_{i=1}^{j-1} \log \frac{P(w_j, w_i)}{P(w_j)P(w_i)}$$

где $(w_1, w_2, \ldots, w_{10})$ – топ-10 слов в рассматриваемой теме t, $P(w_i)$ и $P(w_j)$ – вероятности униграмм w_i и w_j соответственно, а $P(w_j, w_i)$ – вероятность биграммы (w_j, w_i). Итоговая мера согласованности тем вычисляется усреднением $TC\text{-}PMI(t)$ по всем темам t.

Данная метрика показывает очень высокую корреляцию с оценками экспертов [15]. Предложенная метрика рассматривает только первые топ-10 слов в каждой теме, поскольку они, как правило, предоставляют достаточно информации для формирования предмета темы и отличительных черт одной темы от другой. Согласованность тем становится все более широко используемым способом оценивания качества тематических моделей наряду с перплексией. Так, в работе [26] также было показано, что данная метрика очень сильно коррелирует с оценками экспертом. А в работе [27] она просто используется для оценки качества полученных тем.

В соответствии с подходом, изложенным в работе [25], в данной статье вероятности униграмм и биграмм вычисляются путем деления количества документов, в которых встретилась та или иная униграмма или биграмма, на число всех документов в коллекции. Другой вариант вычисления меры согласованности тем на основе логарифма от условной вероятности ($TC\text{-}LCP$), предложенный в работе [25], не рассматривается, поскольку в работе [18] было показано, что этот вариант работает значительно хуже, чем $TC\text{-}PMI$.

4 Добавление биграмм в тематические модели

На первом этапе экспериментов исследовалось, может ли улучшиться качество тематической модели путем добавления в неё биграмм в качестве отдельных элементов словаря. Для этой цели были извлечены все биграммы, встретившиеся в коллекции, с частотностью не меньше 5. Для последующего упорядочения извлечённых биграмм применялись *ассоциативные меры* – математические критерии, определяющие силу связи между составными частями фраз, основываясь на частотах встречаемости отдельных слов и словосочетаний целиком. В экспериментах были использованы следующие 15 ассоциативных мер: *Взаимная Информация (MI)* [28], *Дополненная Взаимная Информация (Дополненная MI)* [29], *Кубическая Взаимная Информация (Кубическая MI)* [30], *Нормализованная Взаимная Информация (Нормализованная MI)* [31], *Настоящая Взаимная Информация (Настоящая MI)*, *Коэффициент Dice (DC)* [32], *Модифицированный Коэффициент Dice (Модифицированный DC)* [33], *T-Score*, *Симметричная Условная Вероятность* [34], *Коэффициент Простого Соответствия*, *Коэффициент Kulczinsky*, *Коэффициент Yula* [30], *Хи-Квадрат*, *Отношение логарифмического правдоподобия* [35] и *Лексическая Связность* [36].

В соответствии с результатами [18] в тематические модели добавлялись топ-1000 биграмм для каждой ассоциативной меры. Так, в каждом эксперименте к словарю в качестве отдельных элементов добавлялись топ-1000 биграмм, и в каждом документе, содержащем любые из добавляемых словосочетаний, из частот образующих их униграмм вычитались частоты биграмм, а сами словосочетания добавлялись в его разреженное представление. Отдельно следует отметить, что во всех экспериментах число топиков фиксировалось равным 100.

Хотя эксперименты были проведены для всех 15 упомянутых выше ассоциативных мер, в таблице 1 представлены только наиболее характерные результаты добавления топ-1000 биграмм наряду с результатом оригинального алгоритма PLSA без добавления биграмм (значения, выделенные полужирным шрифтом, соответствуют улучшению по

одному из критериев).

Ассоциативная мера	Перплексия	TC-PMI
Оригинальный PLSA	1694	86.4
MI	**1683**	79.2
Настоящая MI	2162	**110.7**
Кубическая MI	2000	**95**
DC	1777	**89.6**
Модифицированный DC	2134	**94.1**
T-Score	2189	**104.9**
Лексическая Связность	1928	**101.3**
Хи-Квадрат	1763	**89.6**

Таблица 1: Результаты добавления биграмм в тематическую модель

Как видно, добавление топ-1000 биграмм, упорядоченных по той или иной ассоциативной мере, как правило, приводит к увеличению размера словаря и, следовательно, ухудшению перплексии, в то время как согласованность тем становится лучше. Эти выводы полностью согласуются с результатами, описанными в работе [18]. Однако, используя некоторые ассоциативные меры (например, Взаимную Информацию), можно получить немного лучше перплексию, но чуть хуже согласованность тем, что обусловлено добавлением нестандартных и низкочастотных биграмм.

5 Добавление схожих униграмм и биграмм в тематические модели

5.1 Добавление схожих униграмм в тематические модели

Оригинальные тематические модели (PLSA и LDA) используют модель "мешка слов", предполагающую независимость слов друг от друга. Однако в документах есть много слов, связанных между собой по смыслу – в частности, однокоренные слова, например: *банк – банковский – банкир*, *кредит – кредитный – кредитовать – кредитование* и др. Поэтому на следующем этапе экспериментов исследовалась возможность учета в тематических моделях подобных похожих слов – а именно, слов, начинающихся с одних и тех же букв.

Для данной цели был модифицирован оригинальный алгоритм PLSA. При описании проведённой модификации будет использоваться описание алгоритма PLSA, представленное в работе [37], и следующие обозначения:

- D – коллекция документов;
- T – множество полученных тем;
- W – словарь (множество уникальных слов в коллекции документов D);
- $\Phi = \{\phi_{wt} = p(w|t)\}$ – распределение слов w по темам t;
- $\Theta = \{\theta_{td} = p(t|d)\}$ – распределение тем t по документам d;
- $S = \{S_w\}$ – множество похожих слов, где S_w – множество слов, похожих на w;
- n_{dw} и n_{ds} – частотности слов w и s в документе d;
- $\hat{n}_{wt}$ – оценка частотности слова w в теме t;
- $\hat{n}_{td}$ – оценка частотности темы t в документе d;
- $\hat{n}_t$ – оценка частотности темы t в коллекции документов D.

Псевдокод алгоритма PLSA-SIM представлен в Алгоритме 2. Единственная модификация оригинального алгоритма PLSA касается строчки 6, где в рассмотрение добавляются предварительно вычисленные множества похожих слов (в оригинальном алгоритме данная строчка отсутствует, а в строчке 9 вместо f_{dw} используется n_{dw}). Тем самым вес подобных слов увеличивается в каждом документе коллекции.

Algorithm 2: PLSA-SIM алгоритм: PLSA с похожими словами

Input: коллекция документов D,
количество тем $|T|$,
начальные приближения Φ и Θ,
множества похожих слов S

Output: распределения Φ и Θ

1 **while** *не выполнится критерий остановки* **do**
2 **for** $d \in D$, $w \in W$, $t \in T$ **do**
3 $\hat{n}_{wt} = 0$, $\hat{n}_{td} = 0$, $\hat{n}_t = 0$
4 **for** $d \in D$, $w \in W$ **do**
5 $Z = \sum_t \phi_{wt}\theta_{td}$,
6 $f_{dw} = n_{dw} + \sum_{s \in S_w} n_{ds}$
7 **for** $t \in T$ **do**
8 **if** $\phi_{wt}\theta_{td} > 0$ **then**
9 $\delta = f_{dw}\phi_{wt}\theta_{td}/Z$
10 $\hat{n}_{wt} = \hat{n}_{wt} + \delta$
11 $\hat{n}_{td} = \hat{n}_d + \delta$
12 $\hat{n}_t = \hat{n}_t + \delta$
13 **for** $w \in W$, $t \in T$ **do**
14 $\phi_{wt} = \hat{n}_{wt}/\hat{n}_t$
15 **for** $d \in D$, $t \in T$ **do**
16 $\theta_{td} = \hat{n}_{td}/\hat{n}_t$

Поскольку в русском языке достаточно богатая морфология, а темы в основном задаются именными группами, в качестве потенциальных кандидатов в похожие слова рассматривались только существительные и прилагательные. В таблице 2 представлены результаты добавления похожих слов в тематические модели наряду с оригинальным алгоритмом PLSA (значения, выделенные полужирным шрифтом, соответствуют лучшим значениям по одному из критериев).

Число одинаковых букв	Перплексия	TC-PMI
0 букв (PLSA)	1694	86.4
2 буквы	1852	187.2
3 буквы	1565	432.9
4 буквы	**1434**	**2432.3**
5 букв	1620	**2445.3**
6 букв	1610	1310.85

Таблица 2: Результаты экспериментов по добавлению похожих униграмм в тематическую модель

Как видно, наилучшие результаты показывает модель, рассматривающая в качестве похожих слова, начинающиеся с 4 одинаковых букв. Однако в русском языке есть множество приставок длины в 4 буквы и больше. Учитывая это, был составлен список из 43 наиболее широко использующихся таких приставок (*анти-*, *гипер-*, *пере-* и др.) и введён дополнительный критерий: если слова начинаются на одну и ту же приставку, то они считаются похожими, если следующая буква после приставки также совпадает. Данный критерий позволил еще больше снизить перплексию до **1376** и оставить согласованность тем примерно на лучшем уровне – **2250**. В дальнейших экспериментах, описываемых в данной статье, было решено использовать именно эти 2 критерия.

Следует отметить, что в результате добавления знаний о похожести слов в тематические модели такие слова с большей вероятностью окажутся в топ-10 в полученных темах. Тем самым происходит неявная максимизация меры *TC-PMI*, поскольку похожие слова склонны встречаться в одних и тех же документах. Поэтому было принято решение модифицировать данную метрику для учета не всех топ-10 слов, а только топ-10 непохожих слов в темах (в дальнейшем в статье данная метрика будет обозначаться как *TC-PMI-nSIM*). В таблице 3 подытожены результаты добавления похожих слов в тематические модели с использованием описанных выше критериев и введённой новой метрики:

Алгоритм	Перплексия	TC-PMI-nSIM
Исходный PLSA	1694	78.3
PLSA-SIM	**1376**	**87.8**

Таблица 3: Результаты наилучших способов добавления похожих слов в тематическую модель

Как видно, модифицированная версия алгоритма PLSA-SIM показывает результаты лучше оригинального алгоритма PLSA по обоим целевым метрикам. В таблице 4 представлены топ-5 слов, взятых из двух случайно выбранных тем для оригинального и модифицированного алгоритмов.

PLSA алгоритм		PLSA-SIM алгоритм	
Бумага	Документ	Аудитор	Правый
Ценный	Электронный	Аудиторский	Право
Акция	Форма	Аудитор	Правило
Рынок	Организация	Аудируемый	Акция
Облигация	Подпись	Проверка	Акционер

Таблица 4: Топ-5 слов, взятых из тем, полученных с помощью алгоритмов PLSA и PLSA-SIM

5.2 Добавление схожих биграмм в тематические модели

Для применения подхода, представленного в разделе 5.1 к топ-1000 биграммам, упорядоченными в соответствии с различными ассоциативными мерами, описанными в разделе 4, было решено ввести дополнительный критерий схожести биграмм и униграмм. Биграмма (w_1, w_2) считается похожей на униграмму w_3, если выполнен один из следующих критериев:

- слово w_3 похоже на w_1 или w_2 в соответствии с критериями, описанными в разделе 5.1;
- слово w_3 совпадает с w_1 или w_2 и длина w_3 больше трех букв.

Хотя эксперименты были проведены для всех ассоциативных мер, описанных в разделе 4, в таблице 5 представлены только наиболее характерные результаты интеграции биграмм и добавлению похожести униграмм и биграмм наряду с результатами алгоритмов PLSA и PLSA-SIM (значения, выделенные полужирным шрифтом, соответствуют лучшим значениям по одному из критериев).

Алгоритм	Перплексия	TC-PMI-nSIM
PLSA	1694	78.3
PLSA-SIM	1376	87.8
PLSA-SIM + MI	1411	106.2
PLSA-SIM + Настоящая MI	1204	**177.8**
PLSA-SIM + Кубическая MI	1186	151.7
PLSA-SIM + DC	1288	99
PLSA-SIM + Модифицированный DC	**1163**	156.2
PLSA-SIM + T-Score	1222	171.5
PLSA-SIM + Лексическая связность	1208	125.6
PLSA-SIM + Хи-квадрат	1346	122.9

Таблица 5: Результаты добавления похожих униграмм и биграмм в тематическую модель

Как видно, добавление в тематическую модель похожих униграмм и топ-1000 биграмм, упорядоченных в соответствии с большинством ассоциа-

тивных мер, приводит к улучшению качества получающихся тем по сравнению с алгоритмом PLSA-SIM. В таблице 6 представлены топ-5 униграмм и биграмм, взятых из двух случайно выбранных тем, полученных с помощью алгоритма PLSA-SIM с добавлением топ-1000 биграмм, упорядоченных Модифицированным Коэффициентом Dice (Модифицированным DC), для которого достигаются наилучшее значение перплексии.

Инвестиция	Финансовый рынок
Инвестор	Финансовая система
Инвестирование	Финансовый
Иностранный инвестор	Финансовый институт
Иностранное инвестирование	Финансовый ресурс

Таблица 6: Топ-5 униграмм и биграмм, взятых из тем, полученных с помощью PLSA-SIM с биграммами, упорядоченными Модифицированным DC

6 Итеративный алгоритм для выбора наиболее подходящих биграмм

На последнем этапе экспериментов было сделано предположение, что темы могут сами выбирать себе наиболее подходящие биграммы. Для проверки данной гипотезы был предложен новый итеративный алгоритм выбора биграмм исходя из вида верхушек тем.

При описании предлагаемого алгоритма будут использоваться следующие дополнительные обозначения:

- B – множество всех биграмм в коллекции документов D;
- B_A – множество биграмм, добавленных в тематическую модель;
- S_A – множество потенциальных кандидатов на похожие слова;
- $(u_1^t, \dots, u_{10}^t)$ – топ-10 униграмм в теме t;
- $f(u_1^t, u_2^t)$ – частота биграммы (u_1^t, u_2^t).

Псевдокод предлагаемого алгоритма представлен в Алгоритме 3. На каждой итерации алгоритм добавляет в множество кандидатов в похожие слова топ-10 униграмм из каждой темы. Также в это же множество и в саму тематическую модель добавляются все биграммы, которые могут быть образованы с помощью этих топ-10 униграмм. Было принято решение анализировать только первые топ-10 слов в темах, поскольку одной из целевой метрик является согласованность тем, использующая именно это множество (см. определение метрики в разделе 3). В соответствии с данным алгоритмом темы могут выбирать себе только те биграммы, которые образуются с помощью топ-10 униграмм в темах, а такие биграммы с большей вероятностью могут оказаться наиболее подходящими.

Algorithm 3: Итеративный алгоритм

Input: коллекция документов D,
число тем $|T|$,
множество биграмм B
Output: полученные темы

```
Запуск оригинального PLSA на коллекции документов D для получения тем T
B_A = ∅
while не выполнится критерий остановки do
    S_A = ∅
    for t ∈ T do
        S_A = S_A ∪ {u_1^t, u_2^t, ..., u_10^t}
        for u_i^t, u_j^t ∈ (u_1^t, u_2^t, ..., u_10^t) do
            if (u_i^t, u_j^t) ∈ B and
            f(u_i^t, u_j^t) > f(u_j^t, u_i^t) then
                B_A = B_A ∪ {(u_i^t, u_j^t)}
    S_A = S_A ∪ B_A
    Запуск PLSA-SIM с множеством похожих слов S_A и с множеством биграмм B_A для получения тем T
```

В таблице 7 представлены первые несколько итераций предложенного итеративного алгоритма наряду с результатами оригинального алгоритма PLSA (в таблице обозначен как нулевая итерация).

Итерация	**Перплексия**	**TC-PMI-nSIM**
0 (PLSA)	1694	78.3
1	**936**	**180.5**
2	934	210.2
3	933	230
4	940	**235.8**
5	**931**	193.5

Таблица 7: Результаты итеративного алгоритма построения тематической модели

Как видно, после первой итерации наблюдается существенное улучшение качества получаемых тем по обеим целевым метрикам. Однако на следующих итерациях результаты начинают колебаться вокруг примерно тех же самых уровней перплексии и согласованности тем (с незначительным улучшением последней). Поэтому мы считаем, что согласно результатам первой итерации выбор необходимых биграмм и кандидатов в похожие слова самими темами приводит к наилучшим значениям перплексии и согласованности тем. В таблице 8 приведены топ-5 униграмм и биграмм, взятых из двух случайно выбранных тем, полученных после первой итерации предложенного алгоритма.

Банковский кредит	Ипотечный банк
Банковский сектор	Ипотечный кредит
Кредитование	Ипотечное кредитование
Кредитная система	Жилищное кредитование
Кредит	Ипотека

Таблица 8: Топ-5 униграмм и биграмм, взятых из тем, полученных с помощью итеративного алгоритма построения тематической модели

7 Благодарности

Работа частично поддержана грантом РФФИ 14-07-00383.

8 Заключение

В работе представлены эксперименты по добавлению биграмм в тематические модели. Эксперименты, проведённые на русскоязычных статьях из электронных банковских журналов, показывают, что большинство ассоциативных мер упорядочивает биграммы таким образом, что при добавлении верхушки этих списков в тематические модели ухудшается перплексия и улучшается согласованность тем. Затем в статье предлагается новый алгоритм PLSA-SIM, добавляющий схожесть униграмм и биграмм в тематические модели. Проведённые эксперименты показывают значительное улучшение перплексии и согласованности тем для этого алгоритма. В конце статьи предлагается еще один новый итеративный алгоритм, основанный на идее, что темы сами могут выбирать себе наиболее подходящие биграммы и похожие слова. Эксперименты показывают дальнейшее улучшение качества по обеим целевым метрикам.

Список литературы

[1] D. Blei, A. Ng and M. Jordan. Latent Dirichlet Allocation. Journal of Machine Learning Research, No. 3, pp. 993–1002, 2003.

[2] X. Wei and B. Croft. LDA-based document models for ad-hoc retrieval. In the Proceedings of the 29th International ACM-SIGIR Conference on Research and Development in Information Retrieval, pp. 178–185, 2006.

[3] J. Boyd-Grabber, D. Blei and X. Zhu. A Topic Model for Word Sense Disambiguation. In the Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Processing, pp. 1024–1033, 2007.

[4] D. Wang, S. Zhu, T. Li, and Y. Gong. Multi-Document Summarization using Sentence-based Topic Models. In the Proceedings of the ACL-IJCNLP 2009 Conference Short Papers, pp. 297–300, 2009.

[5] S. Zhou, K. Li, and Y. Liu. Text Categorization Based on Topic Model. International Journal of Computational Intelligence Systems, Vol. 2, No. 4, pp. 398–409, 2009.

[6] L. Bolelli, Ş. Ertekin, C. L. Giles. Topic and Trend Detection in Text Collections Using Latent Dirichlet Allocation. In ECIR Proceedings, Lecture Notes in Computer Science, Vol. 5478, pp. 776–780, 2009.

[7] T. Hyunh, M. Fritz, B. Schiele. Discovery of activity patterns using topic models. In the Proceedings of the 10th international conference on Ubiquitous computing, pp. 10–19, 2008.

[8] T. Hofmann. Probabilistic Latent Semantic Indexing. In the Proceedings of the 22nd Annual International SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57, 1999.

[9] E. Bolshakova, N. Loukachevitch, M. Nokel. Topic Models Can Improve Domain Term Extraction. In ECIR Proceedings, Lecture Notes in Computer Science, Vol. 7814, pp. 684–687, 2013.

[10] M. Nokel, N. Loukachevitch. Application of Topic Models to the Task of Single-Word Term Extraction. In RCDL'2013 Proceedings, pp. 52–60, 2013.

[11] Q. He, K. Chang, E. Lim, A. Banerjee. Keep It Smile with Time: A Reexamination of Probabilistic Topic Detection Models. In the Proceedings of IEEE Transaction Pattern Analysis and Machine Intelligence, vol. 32, issue 10, pp. 1795–1808, 2010.

[12] H. Wallach. Topic Modeling: beyond bag-of-words. In the Proceedings of the 23rd International Conference on Machine Learning, pp. 977–984, 2006.

[13] T. Griffiths, M. Steyvers, and J. Tenenbaum. Topics in semantic representation. Psychological Review, 144, 2, pp. 211–244, 2007.

[14] X. Wang, A. McCallum, and X. Wei. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In the Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, pp. 697–702, 2007.

[15] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin. Automatic evaluation of topic

coherence. In the Proceedings of Human Language Technologies: The 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100-108, 2010.

[16] W. Hu, N. Shimizu, H. Sheng. Modeling chinese documents with topical word-character models. In the Proceedings of the 22nd International Conference on Computational Linguistics, pp. 345–352, 2008.

[17] M. Johnson. PCFGs, topic models, adaptor grammars and learning topical collocations and the structure of proper names. In the Proceedings of the 48th Annual Meeting of the ACL, pp. 1148–1157, 2010.

[18] J. H. Lau, T. Baldwin, and D. Newman. On Collocations and Topic Models. In ACM Transactions on Speech and Language Processing, 10 (3), pp. 1–14, 2013.

[19] D. Andrzejewski, X. Zhu, and M. Craven. Incorporating domain knowledge into topic modeling via Dirichlet Forest priors. In the Proceedings of the 26th Annual International Conference on Machine Learning, pp. 25–32, 2009.

[20] B. Liu. Sentiment Analysis and Opinion Mining. Syntheses Lectures on Human Language Technologies. Morgan & Claypool Publishers. 2012

[21] Z. Zhai, B. Liu, H. Xu, and P. Jia. Grouping Product Features Using Semi-Supervised Learning with Soft-Constraints. In the Proceedings of the 23rd International Conference on Computational Linguistics, pp. 1272–1280, 2010.

[22] A. Daud, J. Li, and F. Muhammad. Knowledge discovery through directed probabilistic topic models: a survey. Frontiers of Computer Science in China, 4(2), pp. 280–301, 2010.

[23] J. Chang, J. Boyd-Grabber, C. Wang, S. Gerrich, and D. Blei. Reading tea leaves: How human interpret topic models. In the Proceedings of the 24th Annual Conference on Neural Information Processing Systems, pp. 288–296, 2009.

[24] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In the Proceedings of the International Conference on Uncertainty in Artificial Intelligence, 2009.

[25] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum. Optimizing semantic coherence in topic models. In the Proceedings of EMNLP'2011, pp. 262–272, 2011.

[26] K. Stevens, P. Kegelmeyer, D. Andrzejewski, D. Butter. Exploring topic coherence over many models and many topics. In the Proceedings of EMNLP-CoNLL'12, pp. 952–961, 2012.

[27] D. Andrzejewski and D. Buttier. Latent topic feedback for information retrieval. In the Proceedings of tthe 17th ACM SIGKDD International Conference on Knowledge discovery and data mining, pp. 600–608, 2011.

[28] K. Church and P. Hanks. Word Association Norms, Mutual Information, and Lexicography. Computational Linguistics, vol. 16, pp. 22–29, 1990.

[29] W. Zhang, T. Yoshida, T. Ho, and X. Tang. Augmented Mutual Information for Multi-Word Term Extraction. International Journal of Innovative Computing, Information and Control, 8(2), pp. 543–554, 2008.

[30] B. Daille. Combined Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering. PhD Dissertation, University of Paris, 1995.

[31] G. Bouma. Normalized Pointwize Mutual Information. In the Proceedings of the Biennal GSCL Conference, pp. 31–40, 2009.

[32] F. Smadja, K. McKeown, and V. Hatzivassiloglou. Translating Collocations for Bilingual Lexicons: A Statistical Approach. Computational Linguistics, 22(1), pp. 1–38, 1996.

[33] M. Kitamura and Y. Matsumoto. Automatic Extraction of Word Sequence Correspondences in Parallel Corpora. In the Proceedings of the 4th Annual Workshop on Very Large Corpora, pp. 79–87, 1996.

[34] J. G. P. Lopes and J. F. Silva. A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units. In the Proceedings of the 6th Meeting on the Mathematics of Language, pp. 369–381, 1999.

[35] T. Dunning. Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, 19(1), 1993.

[36] Y. Park, R. Bird, and B. Boguraev. Automatic Glossary Extraction: Beyond Terminology Identification. In the Proceedings of the 19th International Conference on Computational Linguistics, 2002.

[37] K. Vorontsov and A. Potapenko. EM-like algorithms for probabilistic topic modeling. Machine Learning and Data Analysis, vol. 1(6), pp. 657–686, 2013.

Topic models: taking into account similarity between unigrams and bigrams

Michael Nokel

The paper presents the results of experimental study of integrating word similarity and bigram collocations into topic models. First of all, we analyze a variety of word association measures in order to integrate top-ranked bigrams into topic models. Then we propose a modification of the original algorithm PLSA, which takes into account similar unigrams and bigrams that start with the same beginning. And at the end we present a novel unsupervised iterative algorithm demonstrating how topics can choose the most relevant bigrams. As a target text collection we took articles from various Russian electronic banking magazines. The experiments demonstrate significant improvement of topic models quality for both collections.