УДК 004.056.4

В.А. Минаев[1], И.Д. Королев[2], И.А. Кисленко[2]
([1]МГТУ им. Н.Э. Баумана, [2]Краснодарское высшее военное училище;
e-mail: m1va@yandex.ru)

МЕТОДЫ ВЫЯВЛЕНИЯ ЛАТЕНТНОЙ И НЕГАТИВНОЙ ИНФОРМАЦИИ В ТЕКСТОВЫХ ДОКУМЕНТАХ

Анализируются методы обработки текстовой информации в интересах обеспечения её безопасности. Делается вывод о преимуществах модели латентного размещения Дирихле.

Ключевые слова: текстовый документ, компьютерный анализ, информационная безопасность.

V.A. Minaev, I.D. Korolev, I.A. Kislenko

METHODS FOR LATENT AND NEGATIVE INFORMATION DETECTION IN TEXT DOCUMENTS

Analyzes the methods of textual information processing in order to ensure its safety. It is concluded that the model of Latent Dirichlet Allocation has advantageous characteristics.

Key words: text document, computer analysis, information security.

Статья поступила в редакцию Интернет-журнала 14 сентября 2016 г.

Введение

Сегодня, когда человечество приблизилось к цивилизационной пропасти в связи с накоплением ракетно-ядерного потенциала, способного многократно уничтожить всё живое, западные военные и политические стратеги нашли новый тип оружия – информационное. Оно позволяет путем массированных негативных ***информационно-психологических воздействий (ИПВ)*** на население, различные социальные группы, отдельных индивидов (как правило, лидеров) достигать глобальных целей – от организации массовых неповиновений граждан в тех или иных "нелояльных" государствах до полномасштабных революций. Возникло такое современное явление как "гибридная война". Примеров её реального воплощения немало – Ливия, Египет, Тунис, Сирия, Грузия, Украина, … .

Для противопоставления комплексу специальных операций, проводимых с использованием различных методов, средств и форм распространения информации (через непосредственное общение, путём использования средств массовой информации – печатных средств, радио- и телевещания, изобразительных средств, социальных компьютерных сетей) и инициирующих целенаправленное

изменение поведения объекта воздействия (индивида, группы лиц, массового сознания населения) [7, 9], необходимо создать комплекс эффективных методов выявления, анализа и противодействия ИПВ, как правило, скрытых. Нужно отметить, что такие методы также необходимы для изучения процессов несанкционированного доступа к информации ограниченного доступа и её распространения. Актуальными в этой связи являются современные методы анализа текстовой информации.

Методы компьютерного анализа текстов

Под обработкой текстов на естественном языке понимается использование методов их компьютерного анализа и представления с целью достижения качества, соответствующего обработке "вручную", для последующего применения в различных задачах и приложениях. Одной из актуальных задач автоматической обработки текстов является их кластеризация (выделение групп похожих фрагментов, документов).

Всё чаще применяются статистические тематические методы [1]. Это – современный инструмент анализа текстов. Темы представляются в виде дискретных распределений на множестве слов, а документы – в виде дискретных распределений на множестве тем [1].

Тематические методы осуществляют "нечёткую" кластеризацию слов и документов, означающую, что слово или документ могут быть отнесены сразу к нескольким темам с различными вероятностями. При этом ***синонимы*** с большой вероятностью окажутся в одних и тех же темах, поскольку часто употребляются в рамках одних и тех же документов. В то же время ***омонимы*** (разные по значению, но одинаковые по написанию слова) попадут в различные темы, так как употребляются в различных контекстах [9].

Рассмотрим классическую задачу классификации документов по заданному набору тем. Задача состоит в определении для каждого поступающего в систему документа одной (или нескольких) тем, к которым этот документ относится. Допустим, что в систему не поступает "мусор", то есть каждый из рассматриваемых документов действительно относится хотя бы к одной из заданных тематик.

Тематические методы, как правило, используют метод "мешка слов", в котором каждый документ рассматривается как множество несвязанных между собой слов. Перед выделением тем текст подвергается предобработке, в ходе которой проводится его морфологический анализ с целью определения начальной формы слов и их значений в контексте речи.

Метод обработки машиночитаемых текстов на естественном языке, как правило, базируется на векторно-пространственном методе описания данных (Vector Space Model) [5], предложенном Г. Солтоном в 1975 г.

В рамках указанного метода каждому терму в документе соответствует определенный вес. Таким образом, каждый документ представляется в виде вектора, размерность которого равна общему количеству термов в методе:

сходство документа и темы оценивается как скалярное произведение их некоторых информационных векторов. При этом вес отдельных термов можно вычислять разными способами. Один из них – использование нормализованной частоты его встречаемости в документе:

$$tf(t,d) = \frac{n_i}{\sum_k n_k}, \quad (1)$$

где в числителе n_i – число вхождений i-го слова в документ, а в знаменателе – общее число слов в данном документе.

Вычисленный таким образом вес терма в документе принято обозначать аббревиатурой TF (Term Frequency).

Однако этот подход не учитывает, насколько часто рассматриваемый терм используется во всем массиве документов, то есть так называемую дискриминационную силу терма. Поэтому в случае, когда доступна статистика использования термов во всем массиве документов, более эффективно использовать другой метод, выражаемый формулой:

$$idf(t,D) = \log \frac{|D|}{|(d_i \supset t_i)|}, \quad (2)$$

где $|D|$ – общее количество документов в их массиве, $|(d_i \supset t_i)|$ – количество документов, в которых встречается терм t_i.

Такой метод взвешивания термов указывает не на частоту появления терма в документе, а на величину, обратную количеству документов в массиве, содержащих данный терм (от англ. – inverse document frequency).

Векторно-пространственная метод представления данных обеспечивает системам, построенным на его основе, такие возможности, как:

- создание профессиональных систем и баз знаний;
- повышение уровня компетенции специалистов за счёт получения эффективной возможности направленного поиска и фильтрации текстовых документов;
- автоматическое реферирование текстов документов.

В то же время указанный метод имеет и недостатки, характеризующиеся тем, что:

- матрица терминов-документов очень большая и разреженная;
- близкие по смыслу слова не обязательно встречаются в одних и тех же документах:
- возникает явление синонимии, полисемии, общего шума.

Рассматриваемый метод обработки текстов относится к методам "жёсткой" кластеризации [6], рассматривающим каждый документ как разреженный вектор в пространстве слов большой размерности. Для решения задач обеспечения информационной безопасности векторная модель подходит далеко не всегда, относясь к моделям с жёсткой кластеризацией.

При "мягкой" кластеризации (soft clustering) каждое слово и каждый документ относятся к нескольким темам одновременно с определёнными вероятностями. А семантическое описание слова или документа представляет собой вероятностное распределение на множестве тем. Процесс нахождения этих распределений и называется "тематическим моделированием".

К одним из лучших методов мягкой кластеризации относится латентно-семантический анализ (LSA), который отображает документы и отдельные слова в так называемом "семантическом пространстве", в котором производятся все дальнейшие сравнения.

При этом делаются следующие предположения:

- документы – это набор слов, порядок которых игнорируется. Важно только то, сколько раз то или иное слово встречается в документе;
- семантическое значение документа определяется набором слов, которые, как правило, следуют совместно;
- каждое слово имеет единственное значение.

LSA – это метод обработки информации на естественном языке, анализирующий взаимосвязи между различными массивами документов и терминами в них встречающимися, а также сопоставляющий тематики и документы (термины).

В основе LSA лежат принципы выявления латентных связей изучаемых явлений или объектов. При классификации/кластеризации документов этот метод используется для извлечения контекстно-зависимых значений лексических единиц при помощи статистической обработки очень больших массивов текстов.

В качестве исходной информации LSA использует матрицу термы-на-документы, описывающую набор данных, используемый для обучения системы. Элементы этой матрицы содержат веса, учитывающие частоты использования каждого терма в каждом документе и участие терма во всех документах (TF-IDF).

Наиболее распространенный вариант LSA основан на использовании разложения диагональной матрицы по сингулярным значениям (SVD – Singular Value Decomposition). С помощью SVD-разложения любая матрица раскладывается во множество ортогональных матриц, линейная комбинация которых является достаточно точным приближением к исходной матрице.

Достоинства метода:

- является наилучшим для выявления латентных зависимостей внутри множества документов;
- может быть применён как с обучением, так и без обучения (например, при решении задач кластеризации);
- используются значения матрицы близости, основанной на понятных частотных характеристиках документов и лексических единиц;
- частично снимается полисемия и омонимия.

Недостатки метода:

- значительное снижение скорости вычисления при увеличении объёма входных данных;

- используемый вероятностный метод не всегда соответствует реальности. А именно, предполагается, что слова и документы имеют нормальное распределение, хотя более реально использование распределения Пуассона.

Для устранения указанных недостатков проводится вероятностный латентно-семантический анализ, основанный на мультиномиальном распределении, в частности алгоритм латентного размещения Дирихле (LDA) [2].

В LDA предполагается, что каждое слово в документе порождено некоторой латентной темой, при этом в явном виде используется распределение слов в каждой из них, а также априорное распределение тем в документе. Темы всех слов в документе предполагаются независимыми. В LDA, как и в LSA, документ может соответствовать нескольким темам.

Однако LDA задаёт алгоритм порождения, как слов, так и документов, поэтому появляется дополнительная возможность оценивать вероятности документов вне текстового массива с использованием алгоритма вариационного вывода и семплирования Гиббса.

В отличие от LSA, в LDA число параметров не увеличивается с ростом числа документов в исследуемом массиве. Используемые расширения алгоритма LDA устраняют некоторые его ограничения и улучшают производительность для конкретных задач.

LSA является порождающим алгоритмом только для слов, но не для документов. Алгоритм LDA обходит это ограничение. Основная идея LDA заключается в том, что документы представляются смесью распределений латентных тем, где каждая тема определяется вероятностным распределением на множестве слов. LDA отражает скрытые связи между словами посредством тем, а также позволяет присваивать вероятности новым документам, не входившим в обучающую выборку, используя алгоритм вариационного вывода.

LDA предполагает, что векторы документов $\theta_d = \theta_{td} \in R^T$ и векторы тем $\phi_t = \phi_{wt} \in R^W$ порождаются распределением Дирихле с параметрами $\alpha \in R^T$ и $\beta \in R^W$ (они именуются гиперпараметрами) [8]:

$$\mathrm{Dir}(\theta_d;\alpha) = \frac{(\alpha_0)}{\prod\limits_w (\alpha_t)} \prod_t \theta_{td}^{\alpha_t - 1}; \quad \alpha_t > 0; \quad \alpha_0 = \sum_t \alpha_t; \quad \theta_{td} > 0; \quad \sum_t \theta_{td} = 1;$$

$$\mathrm{Dir}(\phi_t;\alpha) = \frac{(\beta_0)}{\prod\limits_w (\beta_w)} \prod_t \phi_{wt}^{\beta_w - 1}; \quad \beta_0 = \sum_w \beta_w; \quad \phi_{wt} > 0; \quad \sum_t \phi_{wt} = 1. \tag{3}$$

Графически метод LDA можно представить в виде рис. 1.

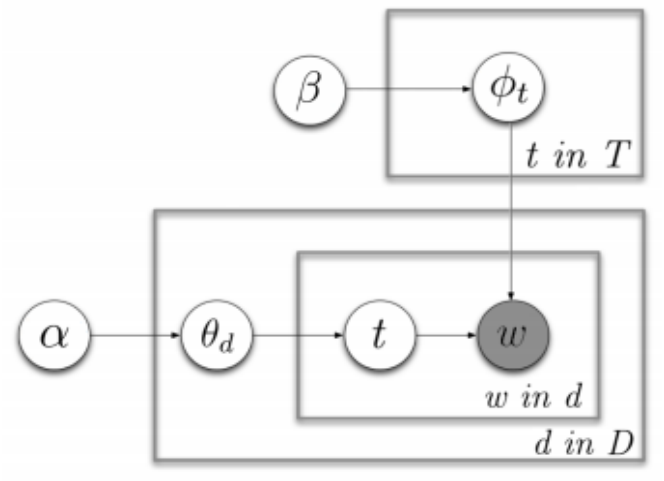

Рис. 1. Графическое представление метода LDA

При описании метода используются следующие обозначения:

w – слово в текстовом массиве;

d – документ в текстовом массиве;

n_{dw}- частотность слова w в документе d;

D – текстовая коллекция;

t – тема;

T – множество выделяемых тем в текстовой коллекции D;

W – множество уникальных слов в текстовой коллекции D(словарь);

$\Phi = \{\phi_{wt}\} = \{P(w|t)\}$ – матрица скрытых распределений тем $P(w|t)$;

$\Theta = \{\theta_{wt}\} = \{P(t|d)\}$ – матрица скрытых распределений документов $P(t|d)$.

Фактически LDA является трехуровневой байесовской сетью, которая порождает документ из смеси тем.

На первом шаге для каждого документа d выбирается случайный вектор θ_d из распределения Дирихле с параметром α (обычно α принимается равным $50/T$).

На втором шаге выбирается тема z_{di} из мультиномиального распределения с параметром θ_d. Наконец, согласно выбранной теме z_{di} выбирается слово w_{di} из распределения Φz_{di}, которое является распределением Дирихле с параметром β (обычно параметр β = 0,1, увеличение β ведёт к более разреженным тематикам).

Поскольку тематический метод зависит от нескольких скрытых переменных, для нахождения оценок максимального правдоподобия параметров Ф и Θ используется EM-алгоритм (Expectation-Maximization) [6]. Это – итеративный алгоритм, каждая итерация которого состоит из двух шагов, повторяющихся до сходимости или до заданного числа итераций:

1. *E-шаг (Expectation-шаг).* На данном шаге вычисляется ожидаемое значение функции правдоподобия, при этом скрытые переменные рассматриваются как наблюдаемые. В рассматриваемой задаче условные вероятности $P(t \mid d, w)$ для всех тем t, документов d и слов w вычисляются через скрытые параметры ϕ_{wt} и θ_{td} по формуле Байеса:

$$P(t \mid d, w) = \frac{P(w,t \mid d)}{P(w \mid d)} = \frac{P(w \mid t)P(t \mid d)}{P(w \mid d)} = \frac{\phi_{wt}\theta_{td}}{\sum_{s \in T} \phi_{ws}\theta_{sd}}. \qquad (4)$$

2. M-шаг (Maximization-шаг). На данном шаге находится оценка максимального правдоподобия. Частотные оценки условных вероятностей вычисляются путём суммирования счётчика $n_{dwt} = n_{dw}P(t \mid d, w)$:

$$\phi_{wt} = \frac{\sum_{d \in D} n_{dwt} + \beta_w}{\sum_{d \in D}\sum_{w \in d} n_{dwt} + \beta_0}; \qquad \theta_{td} = \frac{\sum_{w \in d} n_{dwt} + \alpha_t}{\sum_{w \in W}\sum_{t \in T} n_{dwt} + \alpha_0}. \qquad (5)$$

Наиболее известным критерием является перплексия, используемая для оценки качества различных языковых методов. Перплексия представляет собой меру несоответствия метода $P(w|d)$ словам w, наблюдаемым в документах массива, и определяется через логарифм правдоподобия:

$$\text{Perplexity}(D) = \exp\left(-\frac{1}{n}\sum_{d \in D}\sum_{w \in d} n_{dw} \log P(w \mid d)\right), \qquad 6)$$

где n – число всех рассматриваемых слов в текстовом массиве;

D – множество всех документов в массиве;

n_{dw} – частотность слова w в документе d;

$P(w|d)$ – вероятность появления слова w в документе d.

Чем меньше значение перплексии, тем лучше метод предсказывает появление слов w в документах массива D [6, 10].

Выводы

Таким образом, в результате анализа существующих методов обработки текстов в интересах обеспечения информационной безопасности, следует вывод, что метод LDA наилучшим образом подходит для классификации документов, содержащих латентную информацию, в том числе скрытое ИПВ и данные ограниченного доступа.

В отличие от других методов, с существенным ростом числа документов в обрабатываемом массиве использование LDA позволяет поддерживать производительность автоматизированной системы анализа документов на достаточно высоком уровне.

Литература

1. ***Воронцов К.В., Потапенко А.А.*** Модификации EM-алгоритма для вероятностного тематического моделирования // Машинное обучение и анализ данных. 2013. Т. 1. № 6. С. 657-686.

2. ***Воронцов К.В.*** Вероятностное тематическое моделирование. http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf. Дата обращения: 03.03.2014.

3. ***Минаев В.А., Скрыль С.В., Овчинский А.С., Тростянский С.Н.*** Управление массовым сознанием: современные модели. М.: изд-во РосНОУ, 2013. 200 с.

4. ***Минаев В.А., Дворянкин С.В.*** Обоснование и описание модели динамики информационно-психологических воздействий деструктивного характера в социальных сетях // Безопасность информационных технологий. 2016. № 3. С. 40-52.

5. ***Нокель М.А., Лукашевич Н.В.*** Тематические модели: добавление биграмм и учёт сходства между униграммами и биграммами // Вычислительные методы и программирование. 2015. Т. 16. С. 215-217.

6. ***Носенко С.В., Королев И.Д., Поддубный М.И.*** Способ автоматической классификации формализованных документов в системе электронного документооборота // МПК G06F17/30.

7. ***Asuncion A., Welling M., Smyth P., Teh Y.W.*** On Smoothing and Inference for Topic Models // Proceedings of the International Conference on Uncertainty in Artificial Intelligence. 2009. Pp. 27-34.

8. ***Blei D., Ng A., Jordan M.*** Latent Dirichlet Allocation // Journal of Machine Learning Research. MIT Press, 2003. No. 3. Pp. 993-1002.

9. ***Dempster A.P., Laird N.M., Rubin D.B.*** Maximum Likelihood from Incomplete Data via the EM Algorithm // Journal of the Royal Statistical Society. Series B. 1977. Vol. 39. No 1. Pp. 1-38.

10. ***He Q., Chang K., Lim E., Banerjee A.*** Keep It Smile with Time: A Reexamination of Probabilistic Topic Detection Models // Proceedings of IEEE Transaction Pattern Analysis and Machine Intelligence. 2010. Vol. 32. № 10. Pp. 1795-1808.