УДК 004.44:002.513.5

**Дмитрий Ландэ[1], Цзыцзян Ян[2], Шивэй Чжу[2],
Цзяньпин Го[2], Моцзи Вэй[2]**
[1]Институт проблем регистрации информации НАН Украины
ул. Н. Шпака, 2, 03113 Киев, Украина
[2]Институт информационных исследований Академии наук провинции Шандунь
19, Кейюань Роад, Цзинань, Шандунь, Китай

Автоматическое реферирование китайской правовой информации

Работа посвящена методу автоматического реферирования правовой информации, представленной на китайском языке. Рассмотрена модель реферата и процедура его формирования. Предложено два подхода, а именно: для определения уровня важности предложений — перейти к определению весовых значений отдельных иероглифов, а не слов в тексте документов и рефератов. Также предложено рассмотрение модели документов как сети предложений для выявления наиболее важных предложений по параметрам этой сети. Реализованы и испытаны различные методы автоматического реферирования. Показано применение двух оценок качества реферата без участия экспертов — косинусной меры и дивергенции Дженсена-Шеннона (Jensen-Shan-non). Реферирование на основе предложенной сетевой модели документа оказалось лучшим по критериям косинусной меры и расстояния Дженсена-Шеннона для рефератов, объем которых превышает 2 предложения. Предложенный подход с учетом небольших изменений может использоваться для текстов произвольной тематики, в частности, научно-технической и новостной информации.

Ключевые слова: *автоматическое реферирование, правовая информация, китайский язык, косинусная мера, мера Дженсена-Шеннона.*

Введение

С постановки задач искусственного перевода и автоматического реферирования практически начиналась обработка естественных языков. Первые фундаментальные работы по автоматическому реферированию текстов появились еще в середине прошлого века [1].

Задача связана с решением важнейшей проблемы — сокращением объемов информации, потребляемых человеком, борьбы с информационным шумом. Эта за-

дача весьма актуальна именно сегодня в связи с постоянным ростом информационного пространства. Автоматическое реферирование известно всем пользователям сетевых поисковых систем — в ответ на запрос они получают не только заглавие документа, но и краткое автоматически созданное описание (сниппет). Пользователи мобильных устройств хотят видеть краткое описание статей, прежде чем они переходят к подробному чтению. Лица, принимающие важные управленческие решения, должны ознакамливаться с тысячами документов в сутки, заведомо отметая информационный шум.

В настоящее время существуют сотни промышленных систем автоматического реферирования, например, такие пакеты как Microsoft Office Word AutoSummarize, Mac OS X Summarize, IBM Tivoli Monitoring Summarization and Pruning Agent, Oracle Text, плагины для браузеров Chrome, Mozilla.

Известны многочисленные подходы к автоматическому реферированию, в последнее время все шире применяются нейросетевые технологии, глубинное обучение. Существуют также многочисленные лингвистические подходы, связанные с автоматическим разбором предложений, представленных на различных языках. Традиционный тип систем автоматического реферирования — экстрактивный (квазиреферирование), при котором реферат состоит из отдельных, порой слабо связанных между собой предложений исходного документа. Ему на смену приходит абстрактивный тип реферирования, при котором системы, близкие к системам искусственного интеллекта в сокращенном виде пересказывают содержание исходного документа «своими словами». Однако следует отметить, что сегодня еще практически все промышленные системы автоматического реферирования относятся к экстрактивным системам.

Казалось бы, тема автоматического реферирования текстов уже достаточно изучена, получены основные результаты. Однако и в данной статье речь идет о построении системы автоматического реферирования.

Цель исследования — разработка системы автоматического реферирования. Причин разработки новой системы несколько. Во-первых, решается задача автоматического реферирования правовой информации. А это тексты, которые нельзя в полной мере считать свободными, неструктурированными. Присутствует структура отдельных видов документов, и применение самых лучших универсальных систем реферирования не дает удовлетворительных результатов. Во-вторых, авторы имеют дело с текстами документов, представленными на китайском языке, что существенно сужает круг возможных систем. Кроме того, для обработки китайских текстов, как правило, требуется сегментация слов — в китайском языке слова чаще всего не разделены разделителями. В-третьих, должна быть разработана программа, способная внутри корпоративной системы обрабатывать большие потоки данных с приемлемой производительностью и качеством, встроенная в существующую систему документооборота.

Кроме того, абстрактивный пересказ документов в данном случае неприемлем. Любые «фантазии», вольности пересказа компьютером правовых актов не допустимы. Выход один — разрабатывать некоторый гибридный алгоритм и, соответственно, программу экстрактивного типа, способную учитывать особенности правовых актов КНР, при этом программа должна быть способна обрабатывать отдельные документы, которые объединяются в большие документальные масси-

вы. Эта программа должна быть способна выделять заведомо заданные объекты в помеченных смысловыми маркерами частях документов, выявлять наиболее важные части документов (в том числе и по статистическим критериям), формировать сети предложений и выводить необходимый объем целевой информации в реферат.

Предлагаемый подход

При решении названной проблемы было предложено два подхода, которые можно считать новыми в данной области, а именно: для решения задачи определения уровня важности отдельных частей документов (в нашем случае предложений) было предложено перейти к определению весовых значений отдельных иероглифов, а не слов в тексте документов и рефератов, а также было предложено рассмотрение модели документов как сети предложений для выявления наиболее важных предложений по параметрам этой сети. Вес связей двух предложений в этой сети определяется нормированным некоторым образом весом общих иероглифов, входящих в них.

В рамках традиционного статистического подхода к обработке естественных языков вес предложений обычно вычисляется, исходя из оценочных весов лексических единиц (слов, словосочетаний), входящих в эти предложения [2–5]. В рамках данной работы предлагается в качестве таких элементов для китайского языка использовать отдельные иероглифы.

Переход от рассматриваемых в классической модели слов к иероглифам позволяет избежать относительно сложной процедуры сегментирования слов в тексте, что неизбежно при всех других содержательных методах автоматического анализа китайских текстов. Конечно, данный подход не применим к европейским языкам, где количество различных букв не превышает нескольких десятков. Вместе с тем, для задачи автоматического реферирования китайских текстов предложенный подход дает приемлемые результаты, что будет показано ниже.

Известно, что в китайском языке существует свыше 40 тысяч иероглифов, поэтому каждому из них (пусть и не всегда в полной мере отражающему смысловую единицу) можно приписать весовое значение, рассчитываемое по известным формулам, например, $TF \cdot IDF$ [6].

$TF \cdot IDF$ (от англ. TF — term frequency, IDF — inverse document frequency) — статистическая мера, используемая для оценки важности слова (в данном случае — не слова, а иероглифа) в контексте документа, являющегося частью массива документов. Вес некоторого иероглифа пропорционален количеству его употребления в документе и обратно пропорционален частоте появления этого иероглифа во всех документах массива.

Таким образом, мера $TF \cdot IDF$ зависит от слова t (иероглифа), документа d, всего массива документов D и является произведением двух сомножителей:

$$TF \cdot IDF(t,d,D) = tf(t,d) \times idf(t,D).$$

Здесь выражение $tf(t,d)$ — это отношение числа вхождений некоторого иероглифа к общему количеству иероглифов в документе (к длине документа, фактически). Таким образом, оценивается частота иероглифа в пределах отдельного документа.

Второй сомножитель $idf\left(t, D\right)$ (inverse document frequency — обратная частота документа) — это инверсия частоты, с которой некоторый иероглиф t встречается в документах массива D. Учет idf позволяет уменьшить вес широкоупотребительных иероглифов. Для каждого t в пределах всего массива документов D существует только одно значение idf :

$$idf\left(t, D\right) = \log \frac{\left|D\right|}{\left|\left\{d \in D \mid t \in D\right\}\right|}.$$

Кроме того, в отличие от классических подходов к определению весовых значений предложений, предлагается новая сетевая модель. В рамках этой модели рассматривается ненаправленная сеть, узлами которой выступают отдельные предложения, входящие в документ, между которыми устанавливаются связи в случае наличия у них общих иероглифов. Вес связи между двумя предложениями определяется как сумма весов, общих для этих предложений иероглифов. По этой сети рассчитывается вес каждого предложения как сумма весов связей всех связей, исходящих из соответствующего предложению узла. Естественно, вес предложений затем нормируется, так как длинные предложения без этой процедуры в среднем будут иметь заведомо больший вес. Практика показала, что хорошей нормировкой является деление на логарифм длины соответствующего предложения.

Автоматическое реферирование правовой информации

Процедуры автоматического реферирования экстрактивного класса базируются на определении весовых значений (степени важности) отдельных предложений, которые, в свою очередь, зависят от весов слов. В работе в качестве весовых значений слов использовался классический критерий $TF \cdot IDF$, хотя это не единственный возможный для решения задачи реферирования подход [7]. Традиционно для определения весовых значений предложений использовались два известных алгоритма: в первом случае вес предложения рассматривался как нормированная по длине этого предложения сумма весов входящих в него слов, а во втором — использовался, так называемый, алгоритм симметричного реферирования [8]. В этом случае вес предложения определялся как сумма весов его связей с предыдущим и последующим предложениями.

Кроме того, в данной работе предложен сетевой алгоритм, в котором, в отличие от второго случая, вычисляются связи не только между соседними предложениями, но и между всеми предложениями в тексте документа. Такой подход, конечно, вычислительно более сложный, чем первые два, однако, как показала практика, приводит к лучшим результатам. При этом сложность алгоритма, в случае рассматриваемого подхода реферирования текстов на китайском языке, компенсируется тем, что вместо слов (сегментация которых в данном случае не требуется) рассматриваются лишь отдельные иероглифы.

Итак, приведем основные шаги трех рассматриваемых алгоритмов определения весовых значений предложений.

Шаг 1. Для каждого иероглифа t_i вычисляется значение $DF = df(t_i, d)$ как количество документов d_j из документального массива D, которые содержат данный иероглиф, т.е.:

$$DF := \left| \left\{ d_j \in D : t_i \in d_j \right\} \right|.$$

Шаг 2. Для каждого иероглифа t_i и документа d вычисляется значение $TF = tf\left(t_i, d\right)$ как частоты появления этого термина в документе:

$$TF := \frac{\#\left\{ t_i \in d \right\}}{\left| d \right|}.$$

Затем вычисляется вес иероглифа:

$$w_i = TF \cdot IDF = TF \cdot \log \frac{\left| D \right|}{DF}.$$

Шаг 3. Семгентация предложений, т.е. текст документа разделяется на отдельные предложения p_i, а затем определение их весовых значений wp_i. Введем обозначения: пусть предложение p_i из множества предложений P ($p_i \in P$) состоит из иероглифов $t_{i,k}$ с весом $w_{i,k}$. Запишем в кратком виде суть трех различных алгоритмов.

Шаг 4.1. Алгоритм суммы весов иероглифов ($\sum tf \cdot idf$):

$$wp_i = \frac{1}{\left| p_i \right|} \sum_{k=1}^{\left| p_i \right|} w_{i,k}.$$

Шаг 4.2. Симметричный алгоритм расчета силы связи предложения p_i с ближайшими предложениями (Nearest):

$$wp_i = \frac{1}{\log \left| p_i \right|} \sum_{k=1}^{\left| T \right|} \left(w_{i,k} w_{i-1,k} + w_{i,k} w_{i+1,k} \right),$$

где T — общий состав иероглифов массива. Если иероглиф не присутствует в документе, его вес в нем равен нулю.

Шаг 4.3. Сетевой алгоритм расчета силы связи предложения (Network):

$$wp_i = \frac{1}{\log \left| p_i \right|} \sum_{\substack{j=1, \\ j \neq i}}^{\left| P \right|} \sum_{k=1}^{\left| T \right|} w_{ik} w_{jk} \ .$$

Шаг 5. Вес предложения корректируется в зависимости от его положения в документе. Весовые значения начальных и последних предложений документа искусственно увеличиваются.

Следует отметить, что специфика правовой информации, требования к структуре и объему реферата позволили использовать приведенные выше универсальные подходы к решению частной специальной задачи.

К структуре и объему реферата правового документа (примеры таких документов можно найти на сайте http://www.gov.cn/ в разделе /zhengce) выдвигаются требования, которые нашли свою программную реализацию.

1. Реферат должен начинаться с заголовка документа, приведенного практически без изменений.

2. В реферате отмечается вид документа (объявление «通告», отчет «报告», результаты работы «工作成果», положения «政策» и т.д.).

3. Если в документе обозначена его цель («目的», «奖补目的», «调整目的», «普查的目的和意义» и т.п.), то она также находит отражение в реферате.

4. Если в первом или втором предложении документа обозначены субъекты назначения документов (что также видно по специальным маркерам), то такое предложение также включается в состав реферата.

5. Если в заглавии документа или в обозначении его цели в явном виде присутствуют объекты из числа заранее известных (входящих в таблицу базовых объектов), то эти объекты должны быть выделены в реферате.

6. Если документ относится к типу, не подлежащему дальнейшему реферированию (награды «表彰», объявления о торгах «招标», письма «函» и др.), то реферат считается подготовленным.

7. Из текста документа выбираются все предложения, содержащие выбранные из заглавия и цели объекты. Если таких предложений меньше требуемого числа (заданного заранее или рассчитываемого, исходя из объема документа), то они выводятся в рефераты в той же последовательности, что и в первичном документе. Реферат считается подготовленным.

Если предложений больше требуемого числа, то они взвешиваются в соответствии с приведенным выше алгоритмом (по результатам тестирования выбран сетевой алгоритм). После этого предложения ранжируются по весу и выводятся в реферат в той же последовательности, что и в первичном документе. Реферат считается подготовленным.

В соответствии с представленными требованиями была разработана программа автоматического реферирования правовой информации, представленной на китайском языке. Веб-интерфейс пользователя этой программы приведен на рис. 1.

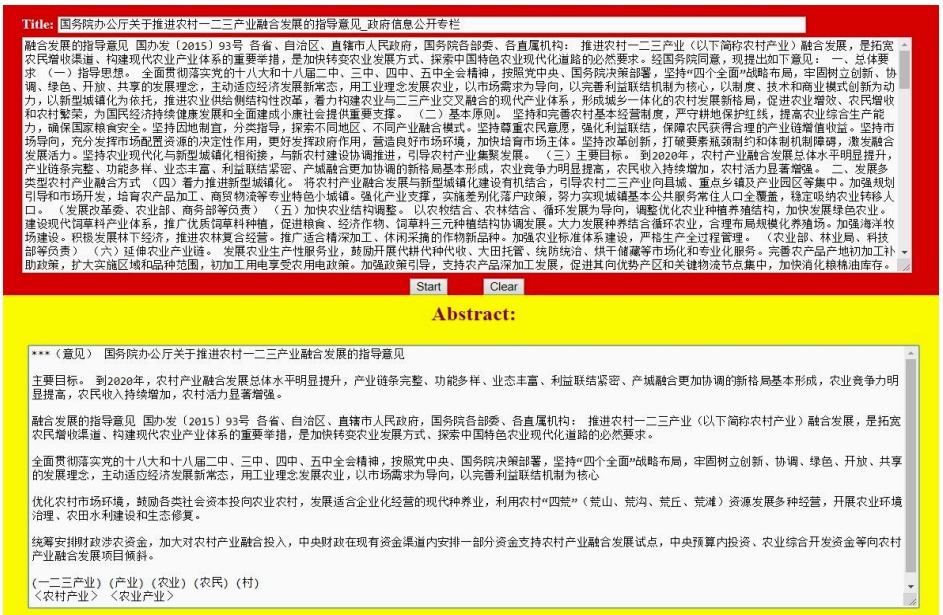


Рис. 1. Веб-интерфейс системы автоматического реферирования

Смежные задачи Text Mining

Автоматическое реферирование текстов — это одна из важных задач технологий глубинного анализа текстов, которая включает еще несколько направлений, таких как извлечение сущностей (Information extraction), построение сетей слов (Language Networks), отражающих особенности предметных областей, кластеризацию (Cluster Analysis).

Предложенный для реферирования алгоритм опирается на некоторое множество заранее подготовленных слов, отражающих основные объекты, представленные в правовых документах (например, «人 口» — население, «产 业» — промышленность, «儿童» — дети и.т.д.).

Вместе с тем, если применить алгоритм сегментации слов, после чего их ранжировать, то легко можно выделить наиболее часто встречающиеся «расширения» стартовых объектов, например, понятие «организация» (组 织) расширить до понятия «международная организация» (国际组织), «общественная организация» (社会组织), а понятие «оборона» (事业) до понятия «народная противовоздушная оборона» (人民防空事业). В результате документам массива правовой информации были поставлены в соответствие основные понятия, которые могут выступать в качестве «ключевых слов», дескрипторов, основ построения моделей предметных областей (Subject Domain).

Как один из видов моделей предметных областей может рассматриваться сеть слов, узлы которой соответствуют отдельным понятиям. Были предложены и реализованы такие простые правила построения этой сети, т.е. правила связи между узлами:

1) все объекты из базового, заранее подготовленного списка, входящие в один документ, связываются связями;

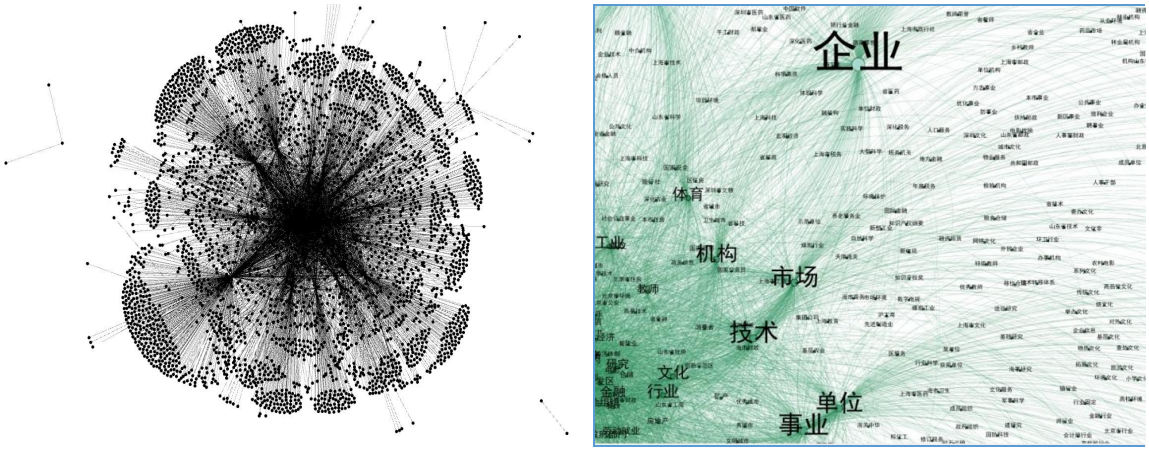
2) если два объекта входят в N разных документов, то сила связи между ними равна N;

3) понятия, являющиеся расширениями понятий из стартового набора, связываются с соответствующими базовыми понятиями.

С помощью программы Gephi (http://gephi.org) [9] построенная сеть была визуализирована (рис. 2), и получены такие параметры построенной сети: количество узлов — 3364 (количество объектов из стартового набора — 220); количество связей — 10167; плотность сети — 0,001; количество связных компонент — 6; средняя длина пути — 3,013; средний коэффициент кластеризации — 0,859.

К топологической особенности построенной сети относится очень большой средний коэффициент кластеризации. Это объясняется, с одой стороны, большим количеством понятий, связанных лишь с порождающим их понятием (отсутствие других соседей), а с другой — сильной связностью объектов из стартового списка. Небольшая средняя длина пути свидетельствует о том, что данная сеть представляет собой «малый мир» (Small World) [10].

С помощью программы Gephi также были получены списки наиболее весомых узлов в соответствии с ранговым критерием PageRank и наибольших посредников по критерию HITS [11] (рис. 3).



Общий вид сети слов　　Фрагмент сети слов

Рис. 2. Сеть слов, отражающая предметную область

Label	PageRank
水利	0.001548
“十一五”	0.00154
扶贫	0.00149
毕业生	0.001408
银行	0.001405
上海市财政	0.001383
邮政	0.001374
林业	0.001352
信息传输	0.001349
城市规划	0.001337
文物	0.001323
医生	0.001315
省财政	0.001242
技术服务	0.001215
山东省财政	0.001194
农民工	0.001126
县域	0.00111
农田	0.001097
电信	0.001072
经济特区	0.001055
科技创新中心	0.001045
试验区	0.001
北京市财政	0.000989
食品药品	0.000964
电影	0.000949
房地产	0.000897
矿产	0.00088
供销	0.000877

PageRank

Label	Hub
残疾人	0.04637
租赁	0.044748
科技创新中心	0.043809
农民工	0.043738
统计	0.04244
电信	0.04112
省财政	0.040688
经济特区	0.039118
山东省财政	0.039081
作业	0.038172
食品药品	0.036646
北京市财政	0.036476
水利	0.036002
试验区	0.035958
电影	0.031812
人工智能	0.031356
娱乐	0.03121
邮政	0.028793
物流业	0.027323
海关	0.026766
社会信用体系建设	0.026739
餐饮	0.02619
深圳市市场	0.02569
干部	0.025645
公共管理	0.025336
金融业	0.025252
食品药品监管	0.025083
经济体制	0.025021

HITS

Рис. 3. Наиболее рейтинговые слова по критериям PageRank и HITS

Приведенный на рис. 2 общий вид сети слов наглядно демонстрирует дальнейшую возможность кластеризации сети, выбора подмножеств — кластеров из слов (понятий). Эта процедура позволяет выделять тематические подмножества в рамках рассматриваемой предметной области.

Методика оценивания результатов

Для оценивания результатов применяется две оценки качества реферата без участия экспертов — косинусная мера и дивергенция Дженсена-Шеннона (Jensen-Shannon), обоснование применения которых приведено в работе [12].

Поясним возможности использования этих подходов. Предполагается, что словарь иероглифов документа d состоит из N элементов $\{t_1, t_2, ..., t_N\}$. Каждому иероглифу соответствует его вес, рассчитанный по правилу $TF \cdot IDF$. Массив этих весовых значений можно представить в виде вектора: $\overline{d} = (w_1, w_2, ..., w_N)$. Соответственно, словарь иероглифов реферата r состоит из подмножества словаря документа, и реферату также можно поставить в соответствие вектор весовых значений: $\overline{r} = (\hat{w}_1, \hat{w}_2, ..., \hat{w}_N)$. При этом дадим естественное определение:

$$\hat{w}_i = \begin{cases} w_i, & if \quad t_i \in r, \\ 0, & if \quad t_i \notin r. \end{cases}$$

Известно, что скалярное произведение двух ненулевых векторов в евклидовом пространстве A и B определяется формулой

$$\overline{A} \cdot \overline{B} = \left\|\overline{A}\right\| \left\|\overline{B}\right\| \cos\theta .$$

Здесь θ — угол между рассматриваемыми векторами. Естественно, если направление векторов совпадает, значение θ становится равным нулю (соответственно $\cos\theta = 1$). То есть, чем ближе $\cos\theta$ к единице, тем направление векторов ближе, что легко содержательно интерпретируется для случая документа и его реферата (краткой аннотации). Принято функцию близости между векторами A и B обозначать как $Sim(\overline{A}, \overline{B})$ (от английского слова Similarity). В случае изучения косинусной меры близости имеем

$$Sim\left(\overline{A}, \overline{B}\right) = \cos\theta = \frac{\overline{A} \cdot \overline{B}}{\left\|\overline{A}\right\| \left\|\overline{B}\right\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}},$$

где A_i и B_i — компоненты векторов $\overline{A}$ и $\overline{B}$, соответственно.

В соответствии с определением косинусной меры для вычисления близости документа и реферата можно воспользоваться формулой:

$$Sim(d, r) = \frac{\sum_{i=1}^{N} w_i \hat{w}_i}{\sqrt{\sum_{i=1}^{n} w_i^2} \sqrt{\sum_{i=1}^{n} \hat{w}_i^2}}.$$

Исходя из того, что $\sum_{i=1}^{N} w_i \hat{w}_i = \sum_{i=1}^{N} \hat{w}_i \hat{w}_i$, получаем формулу, которая используется при практических расчетах:

$$Sim(d,r) = \frac{\sum_{i=1}^{N} \hat{w}_i \hat{w}_i}{\sqrt{\sum_{i=1}^{n} w_i^2} \sqrt{\sum_{i=1}^{n} \hat{w}_i^2}} = \sqrt{\frac{\sum_{i=1}^{N} \hat{w}_i^2}{\sum_{i=1}^{N} w_i^2}}.$$

Другой, используемый нами критерий формального выявления степени близости — дивергенция Дженсена-Шеннона, которая базируется на формализме теории информации и математической статистики, в частности, на относительной энтропии (relative entropy) Кульбака-Лейблера (Kullback-Leibler) [13, 14].

Энтропия Кульбака-Лейблера в общем случае определяется как неотрицательнозначный функционал, являющийся несимметричной мерой удаленности друг от друга двух вероятностных распределений, определенных на общем пространстве элементарных событий.

Расхождение распределения Q относительно P обозначается $D(P\|Q)$. Распределение Q часто служит приближением распределения P. Данная мера расстояния в теории информации также интерпретируется как величина потерь информации при замене истинного распределения P на распределение Q. Значение функционала можно понимать как количество неучтенной информации распределения Q, если оно было использовано для приближения P.

Для дискретных вероятностных распределений $P\{p_1, p_2, ..., p_n\}$ и $Q\{q_1, q_2, ..., q_n\}$ энтропия Кульбака-Лейблера определяется следующим образом:

$$D(P\|Q) = \sum_{i=1}^{n} \log \frac{p_i}{q_i} p_i.$$

Энтропию Кульбака-Лейблера, содержательно близкую к понятию расстояния, можно было бы назвать метрикой в пространстве вероятностных распределений, но это было бы некорректно, так как оно не симметрично $D(P\|Q) \neq D(Q\|P)$ и не удовлетворяет неравенству треугольника.

В дальнейшем мы будем использовать дивергенцию Дженсена-Шеннона (JSD), которая базируется на энтропии Кульбака-Лейблера, но при этом является метрикой [15, 16], поэтому она еще называется «расстоянием Дженсена-Шеннона» [17–19].

Дивергенция Дженсена-Шеннона определяется следующим образом:

$$JSD(P\|Q) = \frac{1}{2}\left(D(P\|M) + D(Q\|M)\right),$$

где $M = \frac{1}{2}(P+Q)$.

В случае применения расстояния Дженсена-Шеннона к задаче оценки качества рефератов оценивается количество потерянной информации в реферате по

сравнению с исходным документом. Так же, как и в косинусной мере, предполагается, что документу d соответствует вектор весов иероглифов $\bar{d} = (w_1, w_2, ..., w_N)$, а реферату r — вектор весовых значений: $\bar{r} = (\hat{w}_1, \hat{w}_2, ..., \hat{w}_N)$. Используемый в методе Дженсена-Шеннона «средний» вектор представляется в следующем виде:

$$\bar{M} = \frac{1}{2}\left(\bar{d} + \bar{r}\right).$$

Соответственно:

$$JSD = \frac{1}{2}\left(D(\bar{d}\,\|\bar{M}) + D(\bar{r}\,\|\bar{M})\right) = \frac{1}{2}\left(\sum_{i=1}^{N}\log\left(\frac{w_i}{\frac{1}{2}(w_i + \hat{w}_i)}\right)w_i + \sum_{i=1}^{N}\log\left(\frac{\hat{w}_i}{\frac{1}{2}(w_i + \hat{w}_i)}\right)\hat{w}_i\right).$$

Рассмотрим приведенные суммы на двух областях значения индексов *i*: 1-я область, где иероглифы документа и реферата совпадают, и 2-я, где не совпадают, т.е., где $\hat{w}_i = 0$:

$$JSD = JSD_1 + JSD_2\,.$$

В первой области, очевидно:

$$JSD_1 = \frac{1}{2}\sum_{i=1}^{N}\log\left(\frac{w_i}{\frac{1}{2}(w_i + w_i)}\right)w_i + \frac{1}{2}\sum_{i=1}^{N}\log\left(\frac{w_i}{\frac{1}{2}(w_i + w_i)}\right)w_i = 0.$$

Во второй области, соответственно:

$$JSD_1 = \frac{1}{2}\sum_{i=1}^{N}\log\left(\frac{w_i}{\frac{1}{2}w_i}\right)w_i + \frac{1}{2}\sum_{i=1}^{N}\log\left(\frac{\hat{w}_i}{\frac{1}{2}(w_i)}\right)\cdot\hat{w}_i \to \frac{1}{2}\sum_{i=1}^{N}w_i.$$

Строго говоря, второе слагаемое в последней формуле не корректно (можно рассмотреть предел выражений под знаком суммы при $\hat{w}_i \to 0$), но вместе с тем, можно сделать достаточно очевидный вывод, что мера Дженсена-Шеннона соответствует потере информации при реферировании и пропорциональна суммарному весу слов (в нашем случае — иероглифов), входящих в документ, но отсутствующих в реферате.

Сравнение методов

При реферировании была реализована новая идея определения весовых значений предложений на основе весов отдельных иероглифов, а не слов, как это общепринято. Поэтому качество реферирования проверяется не только, исходя из учета весов отдельных иероглифов, но и с учетом весов целых слов, входящих в документы и рефераты, чтобы убедиться, что предложенный подход удовлетворителен и по критериям традиционных систем реферирования. Естественно, для

этого пришлось выполнить затратную по ресурсам процедуру сегментации слов [20]. Следует отметить, что данная процедура выполнялась исключительно для проверки качества алгоритмов реферирования, и она не входит в состав самих этих алгоритмов.

Исследования проводились на реальном массиве правовой информации Китайской народной республики объемом 10 тысяч документов.

На рис. 4–7 показаны результаты проведенных исследований. На рис. 4 и 6 приведены результаты, когда моделям документов и рефератов соответствовали векторы, элементы которых — весовые значения отдельных иероглифов из текста документа по $TF \cdot IDF$. На рис. 5 и 7 — результаты, когда элементами векторов соответствуют весовым значениям слов, сегментированных из текстов документов и рефератов. На рис. 4 и 6 приведены результаты в соответствии с косинусной мерой оценки близости документа и реферата, а на рис. 5 и 7 — в соответствии с расстоянием Дженсена-Шеннона.

По горизонтальной оси на всех рисунках отмечено количество предложений, входящих в реферат. По вертикальной оси приведены усредненные по всему массиву документов значения соответствующих критериев. Следует отметить, что во всех примерах в качестве первого предложения реферата входит заглавие документа, поэтому значения с аргументом 1 для всех четырех видов алгоритмов (Σ *tf·idf*, Nearest, Network, Random) совпадают. Как видно, для сравнения к трем приведенным выше алгоритмам добавлен метод Random — составление реферата из случайных предложений текста (кроме первого предложения — заглавия).

	1	2	3	4	5	6	7	8	9	10
Σtf-idf	0.851	0.936	0.951	0.958	0.967	0.97	0.974	0.976	0.978	0.98
Network	0.851	0.9	0.936	0.951	0.962	0.967	0.97	0.973	0.975	0.976
Nearest	0.851	0.89	0.93	0.944	0.954	0.961	0.965	0.97	0.972	0.974
Random	0.851	0.888	0.907	0.92	0.942	0.952	0.955	0.957	0.964	0.966

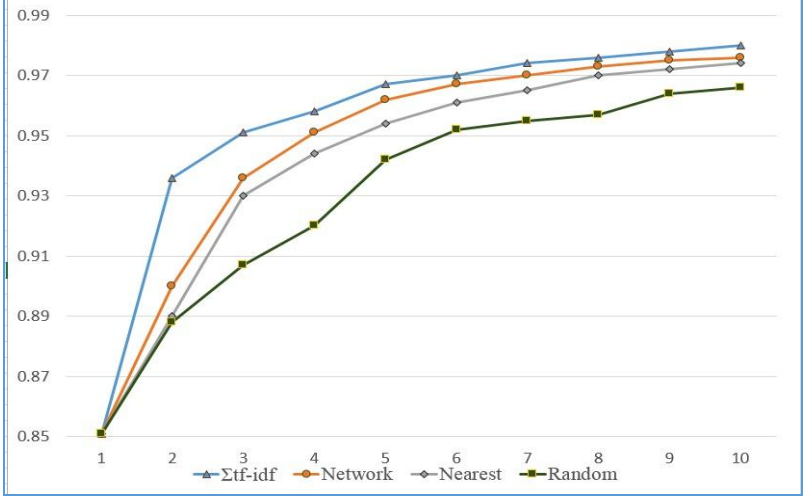


Рис. 4. Косинусная мера близости документа и реферата — учет весовых значений отдельных иероглифов

	1	2	3	4	5	6	7	8	9	10
Σtf-idf	0.392	0.331	0.313	0.293	0.272	0.256	0.241	0.23	0.22	0.212
Network	0.392	0.324	0.274	0.247	0.223	0.204	0.189	0.177	0.167	0.159
Nearest	0.392	0.342	0.293	0.262	0.242	0.225	0.21	0.2	0.189	0.182
Random	0.392	0.353	0.326	0.303	0.283	0.261	0.246	0.23	0.216	0.208

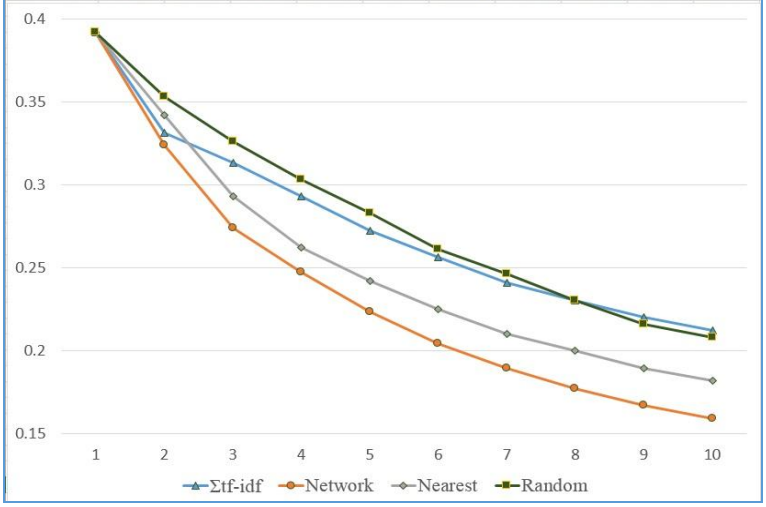


Рис. 5. Мера Дженсена-Шеннона потери информации при реферировании — учет весовых значений отдельных иероглифов

	1	2	3	4	5	6	7	8	9	10
Σtf-idf	0.422	0.715	0.807	0.835	0.853	0.869	0.879	0.888	0.893	0.9
Network	0.422	0.715	0.788	0.836	0.861	0.882	0.894	0.903	0.91	0.915
Nearest	0.422	0.715	0.775	0.824	0.85	0.868	0.88	0.891	0.898	0.905
Random	0.422	0.576	0.663	0.735	0.769	0.789	0.811	0.823	0.839	0.849

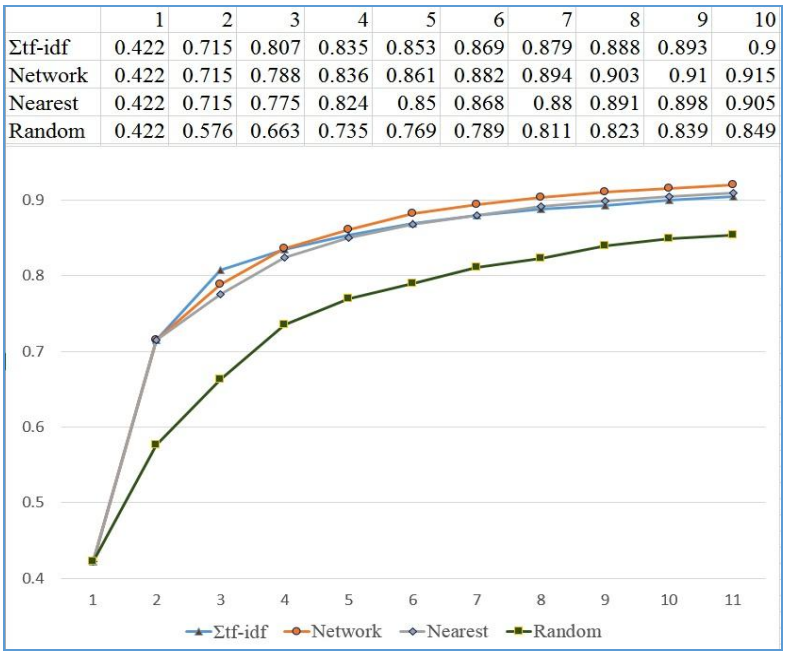


Рис. 6. Косинусная мера близости текста и реферата — учет весовых значений отдельных слов

	1	2	3	4	5	6	7	8	9	10
Σtf-idf	0.419	0.365	0.35	0.332	0.314	0.3	0.288	0.279	0.271	0.264
Network	0.419	0.367	0.324	0.301	0.278	0.261	0.246	0.234	0.225	0.216
Nearest	0.419	0.379	0.335	0.307	0.289	0.275	0.262	0.253	0.243	0.235
Random	0.419	0.391	0.367	0.346	0.322	0.314	0.296	0.286	0.276	0.268

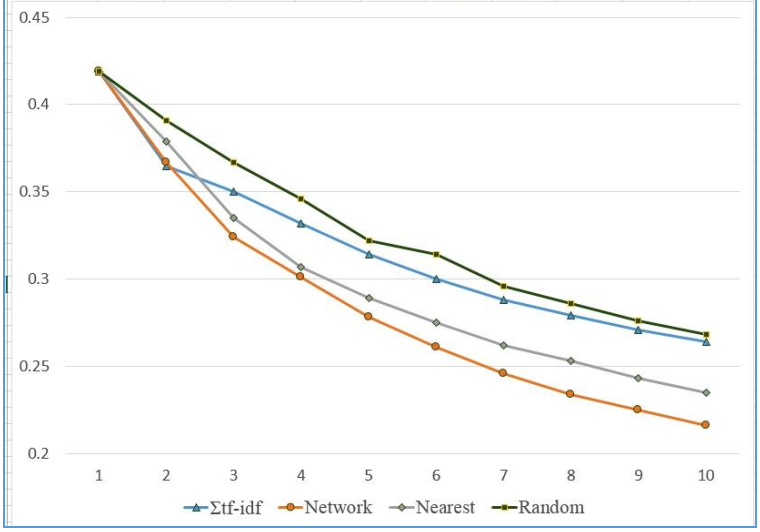


Рис. 7. Мера Дженсена-Шеннона потери информации при реферировании — учет весовых значений отдельных слов

Результаты испытаний позволяют резюмировать следующее.

Предложенные подходы приводят к результатам, качество которых не ниже, представленных на известной конференции по анализу текстов TAC [Lois, 2008].

Если по критерию косинусной меры близости документа и реферата при учете весовых значений отдельных иероглифов наилучшие результаты показал метод $\sum tf \cdot idf$ (что естественно, по сумме $TF \cdot IDF$ определялся вес предложений, при этом самые весомые входили в реферат), то по тем же критериям, при учете отдельных слов естественного языка наилучшим оказался предложенный сетевой метод.

Выводы

Представлена новая гибридная методика автоматического реферирования, охватывающая статистические и маркерные методы, а также учет расположения предложений в тексте документа. Предложенная модель реферата отражает информационную потребность заказчиков при работе с правовой информацией.

Реализован подход к определению весовых значений отдельных иероглифов, а не сегментированных слов в тексте документов и рефератов. Данная методика позволяет избегать затратной процедуры сегментирования слов, необходимой для других содержательных методов обработки китайского языка.

Реализованы и испытаны различные методы автоматического реферирования. Реферирование на основе предложенной сетевой модели документа оказалось лучшим по критериям косинусной меры и расстояния Дженсена-Шеннона для рефератов, объем которых превышает 2 предложения.

Предложенный подход, с учетом небольших изменений в маркерах-шаблонах, может использоваться не только для правовых документов, но и для текстов произвольной тематики, в частности, научно-технической и новостной информации.

1. Luhn Hans Peter. The automatic creation of literature abstracts. *IBM Journal of research and development*. 1958. N 2. P. 159–165.

2. Zhang C. Automatic Keyword Extraction from Documents using Conditional Random Fields. *Journal of Computational Information Systems*. 2008. **4**(3). P. 1169–1180.

3. Ramos J. Using tf-idf to determine word relevance in document queries. Proceedings of the first instructional conference on machine learning, 2003. P. 1–4.

4. Bhart, Santosh Kumar, Babu Korra Sathya, Pradhan, Anima. Automatic Keyword Extraction for Text Summarization in Multi-document e-Newspapers Articles. *European Journal of Advances in Engineering and Technology*. 2017. **4**(6). P. 410–427.

5. Chien L.-F. Pat-tree-based keyword extraction for Chinese information retrieval. ACM SIGIR Forum. 31, ACM, 1997. P. 50–58.

6. Salton G., Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*. 1998. **24**(5). P. 513–523.

7. Lande D.V., Snarskii A.A, Yagunova E.V., Pronoza E.V. The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text. 12th Mexican International Conference on Artificial Intelligence. 2013. P. 209–215. DOI: 10.1109/MICAI.2013.33.

8. Yatsko V.A. Symmetric Summarization: Thematic Foundations and Methods. Nauchno-Tekh. Inf., 2002. Ser. 2. N 5. P. 18–28.

9. Cherven Ken. Network Graph Analysis and Visualization with Gephi. Packt Publishing, 2013. ISBN: 9781783280131.

10. Kleinberg J. Navigation in a small world. *Nature*. 2000. **406**(6798). 845 p. DOI: 10.1038/35022643.

11. Langville Amy N., Meyer Carl D. Google's PageRank and beyond: the science of search engine rankings. Princeton university press, 2011. ISBN: 9780691152660.

12. Louis Annie, Nenkova Ani. Automatic Summary Evaluation without Human Models. In First Text Analysis Conference (TAC'08). Gaithersburg, MD, Etats-Unis, 17–19 November, 2008.

13. Kullback S. Information Theory and Statistics. John Wiley & Sons, 1959. Republished by Dover Publications in 1968; reprinted in 1978: ISBN 0-8446-5625-9.

14. Kullback S., Leibler R.A. On information and sufficiency. *Annals of Mathematical Statistics*, 1951. **22**(1). P. 79–86. DOI: 10.1214/aoms/1177729694.

15. Schütze Hinrich, Manning Christopher D. Foundations of Statistical Natural Language Processing. Cambridge, Mass: MIT Press, 1999. 304 p. ISBN 0-262-13360-1.

16. Dagan Ido, Lillian Lee, Fernando Pereira. Similarity-Based Methods For Word Sense Disambiguation. Proc. Of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics. 1997. P. 56–63. DOI: 10.3115/979617.979625.

17. Endres D.M., Schindelin J.E. A new metric for probability distributions. IEEE Trans. Inf. Theory, 2003. **49**(7). P. 1858–1860. DOI: 10.1109/TIT.2003.813506.

18. Fuglede Bent, Topse Flemming. Jensen-Shannon divergence and Hilbert space embedding. Proc. Of International Symposium on Information Theory, ISIT 2004. P. 31.

19. Lin J. Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory, 1991. **37**(1). P. 145–151. DOI:10.1109/18.61115.

20. Berezin Boris A., Lande Dmitry V., Pavlenko Oleh Y. Development, Evaluation and Usage of Word Segmentation Algorithm for National Internet Resources Monitoring Systems. CEUR Workshop Proceedings, 2017. Selected Papers of the XVII International Scientific and Practical Conference on Information Technologies and Security (ITS 2017). 2017. P. 16–22.

Поступила в редакцию 08.09.2018