

Учреждение образования
«Белорусский государственный университет
информатики и радиоэлектроники»



**Информационные технологии и системы 2015
(ИТС 2015)**

Материалы международной научной конференции
(БГУИР, Минск, Беларусь, 28 октября 2015)

**Information Technologies and Systems 2015
(ITS 2015)**

Proceeding of The International Conference
(BSUIR, Minsk, Belarus, 28th October 2015)

Минск БГУИР 2015

УДК 004
ББК 32.973.202
И74

Редакционная коллегия:

Л. Ю. Шилин (главный редактор), Д. П. Кукин, А. А. Иванюк,
А. В. Марков, В. В. Голенков, А. А. Навроцкий, А. Ф. Трофимович,
Н. С. Жилач, С. В. Колосов, А. Б. Гуринович (ответственный секретарь)

И74 Информационные технологии и системы 2015 (ИТС 2015)
: материалы международной научной конференции (БГУИР, Минск, Беларусь, 28 октября 2015)=Information Technologies and Systems 2015 (ITS 2015) : Proceeding of the International Conference (BSUIR, Minsk, Belarus, 28th October 2015) / редкол. : Л. Ю. Шилин [и др.]. – Минск : БГУИР, 2015. – 324с.
ISBN 978-985-543-171-9.

Сборник включает прошедшие рецензирование доклады международной научной конференции «Информационные технологии и системы 2015» (ИТС 2015).

Предназначен для преподавателей высших учебных заведений, научных сотрудников, студентов, аспирантов, магистрантов, а также для специалистов предприятий в сфере IT-технологий.

Материалы сборника одобрены организационным комитетом и печатаются в авторской редакции.

УДК 004
ББК 32.973.202

ISBN 978-985-543-171-9

© УО «Белорусский государственный университет информатики и радиоэлектроники», 2015

РЕФЕРИРОВАНИЕ ОБРАЗОВАТЕЛЬНЫХ ТЕКСТОВ ЧЕРЕЗ КЛАСТЕРИЗАЦИЮ И РАНЖИРОВАНИЕ ПРЕДЛОЖЕНИЙ

Л. М. Огнивчук

Кафедра информационных технологий и математических дисциплин, Киевский университет имени

Бориса Гринченка

Киев, Украина

E-mail: l.ohnivchuk@kubg.edu.ua

Предлагается метод реферирования текстов образовательной сферы деятельности на основе схемы содержательных аспектов соответствующего класса документов путем предварительной кластеризации предложений. Специфика этого подхода заключается в том, что образованный реферат будет содержать основной смысл различных тем наиболее полно и с меньшей избыточностью.

ВВЕДЕНИЕ

Важным фактором при автоматическом реферировании текста является учет его тематической структуры. Тематическая структура содержит в себе дополнительную информацию о внутреннем устройстве документа, которая может быть использована для улучшения автоматических операций над текстовыми данными [1].

Идея метода, предложенного в данной работе, заключается в том, чтобы сначала найти тематические разделы набора документов, то есть группы предложений, относящихся к одной под теме. После кластеризации, с учетом степени значимости кластеров, применяя алгоритм ранжирования, вытягиваются информативные предложения.

При ручном реферировании широко используется поаспектный метод реферирования, который заключается в выборе содержательных аспектов в первичном документе и создании на их основе вторичного документа [2]. Аспекты текстов едины для различных отраслей знаний, хотя и отличаются содержанием и формой.

Наибольший интерес для нас представляют тексты образовательной сферы деятельности [3]. Тексты образовательной сферы деятельности можно условно разделить на научно-исследовательские разработки и тексты психолого-педагогической тематики.

Задача заключается в нахождении метода, который позволит автоматически выделять тематическую структуру входящего документа образовательной сферы деятельности с использованием схем содержательных аспектов.

В структуре научных текстов выделяют формальные текстовые признаки - устойчивые языковые высказывания, своеобразные речевые клише, штампы, позволяющие различать отдельные аспекты содержания в тексте, проследить развитие авторской мысли в тексте. К таким формальным текстовым признакам относятся маркеры [2]. Каждый аспект имеет свой специфический набор маркеров, причем в текстах

разных областей знаний маркеры одних и тех же аспектов не имеют существенных различий.

В своей работе референты используют схемы содержательных аспектов. Не в каждом тексте можно найти все аспекты схем, но большинство из них присущи почти всем научным документам. Поэтому при решении задачи кластеризации предложений текстов образовательной сферы деятельности с применением метода k-средних с целью их дальнейшего ранжирования, считаем, что первоначальное число кластеров совпадает с количеством содержательных аспектов, а начальное число центров кластеров будем определять на основе коэффициентов сходства предложений документа к его содержательным аспектам.

КЛАСТЕРИЗАЦИЯ ПРЕДЛОЖЕНИЙ ДОКУМЕНТА

Обозначим $S = \{s_1, s_2, \dots, s_{N_{s,d}}\}$ — множество предложений документа d , где $N_{s,d}$ — общее количество предложений в документе d . $A = \{A_1, A_2, \dots, A_{N_A}\}$ — множество смысловых аспектов, где N_A — общее количество смысловых аспектов. $A_j \cap A_l = \emptyset$ при $j \neq l$, $j = 1, \dots, N_A$, $l = 1, \dots, N_A$. $\cup_{k=1}^{N_A} A_k = S$. $O = \{O_1, O_2, \dots, O_{N_A}\}$ — вектор центров смысловых аспектов.

Задача кластеризации заключается в том, что каждой паре $(s_i, A_j) \in S \times A$ необходимо поставить в соответствие значение $\{0, 1\}$.

В соответствии с алгоритмом кластеризации по методу k-средних задаем число кластеров, которое на первоначальном этапе совпадает с количеством смысловых аспектов в соответствии со схемой смысловых аспектов, то есть $k = N_A$.

Далее выбираем k центров кластеризации следующим образом

$$O_j = \max_i k_{i,j}, \quad (1)$$

где $k_{i,j}$ — степень принадлежности предложения s_i к смысловому аспекту A_j , $i = 1, \dots, N_{s,d}$, $j = 1, \dots, k$.

Для определения степени принадлежности предложения s_i к смысловому аспекту A_j введем следующие обозначение.

Обозначим $s_i = \{t_{11}, t_{12}, \dots, t_{1N_{t_1, s_i}}\}$ – множество однословных термов предложения s_i , N_{t_1, s_i} – общее количество однословных термов предложения s_i . Аналогично $s_i = \{t_{21}, t_{22}, \dots, t_{2N_{t_2, s_i}}\}$ – множество двухсловных термов предложения s_i , N_{t_2, s_i} – общее количество двухсловных термов предложения s_i . $s_i = \{t_{31}, t_{32}, \dots, t_{3N_{t_3, s_i}}\}$ – множество трехсловных термов предложения s_i , N_{t_3, s_i} – общее количество трюхсловных термов предложения s_i .

Обозначим $A_j = \{m_{11}, m_{12}, \dots, m_{1N_{m_1, A_j}}\}$ – множество однословных маркеров смыслового аспекта A_j , $i = 1, \dots, N_{s, d}$, N_{m_1, A_j} – общее количество однословных маркеров смыслового аспекта A_j . Обозначение для двухсловных и трехсловных маркеров аналогично.

Тогда количество однословных маркеров смыслового аспекта A_j , которые появляются в предложении s_i находим по формуле

$$E_{1ij} = \sum_{r=1}^{N_{m_1, A_j}} \sum_{\tau=1}^{N_{t_1, s_i}} |m_{1r, A_j} \cap t_{1\tau, s_i}|.$$

Формулы для двухсловных и трехсловных маркеров смыслового аспекта A_j аналогичные.

Обозначим $K_{1ij}, K_{2ij}, K_{3ij}$ – степени принадлежности соответственно одно- двух- и трюхсловных маркеров смыслового аспекта A_j к предложению s_i , которые определяются как простое отношение количества маркеров смыслового аспекта A_j , появляющиеся в предложении s_i к общему числу соответствующих маркеров в смысловом аспекте A_j

$$K_{1ij} = \frac{E_{1ij}}{N_{m_1, A_j}}, K_{2ij} = \frac{E_{2ij}}{N_{m_2, A_j}}, K_{3ij} = \frac{E_{3ij}}{N_{m_3, A_j}}. \quad (2)$$

Учитывая формулы (2) степень принадлежности предложения s_i к смысловому аспекту A_j для формулы (1) будем определять по формуле

$$K_{ij} = \frac{\xi_1 K_{1ij} + \xi_2 K_{2ij} + \xi_3 K_{3ij}}{3}, \quad (3)$$

где ξ_1, ξ_2, ξ_3 – весовые коэффициенты одно-двух- и трех словных маркеров соответственно. Сходство между двумя предложениями s_i и s_ζ на основе степени косинусов определяется как:

$$\begin{aligned} \text{sim}(s_i, s_\zeta) &= \cos(s_i, s_\zeta) = \\ &= \frac{\sum_{\chi=1}^{N_{w, d}} (\omega_{i\chi} \omega_{\zeta\chi})}{\sqrt{\sum_{\chi=1}^{N_{w, d}} (\omega_{i\chi}^2)} \sqrt{\sum_{\chi=1}^{N_{w, d}} (\omega_{\zeta\chi}^2)}}. \end{aligned} \quad (4)$$

С оптимистической точки зрения, если $s_i \in A_j$, $s_\zeta \in A_l$ при $j \neq l, i \neq \zeta$, то мера близости между ними должна быть минимальной, то есть

$$\text{sim}(s_i, s_\zeta) \rightarrow \min, \quad (5)$$

а расстояние между предложением и соответствующим смысловым аспектом к которому оно

принадлежит должна быть максимальным, то есть

$$\text{sim}(s_i, O_j) \rightarrow \max. \quad (6)$$

На первом этапе кластеризации центр кластера определяется по формуле (1), а для каждой последующей итерации по формуле

$$O_j = \frac{\sum_{i=1}^{N_{s, d}} u_{ij} s_i}{\sum_{i=1}^{N_{s, d}} u_{ij}} \quad (7)$$

где u_{ij} – степень принадлежности предложения s_i к кластеру A_j . При этом $\sum_{j=1}^k u_{ij} = 1, \forall i = 1, \dots, N_{s, d}$. $\sum_{i=1}^{N_{s, d}} u_{ij} > 0, \forall j = 1, \dots, k$. $u_{ij} = \{0, 1\}, \forall i = 1, \dots, N_{s, d}, \forall j = 1, \dots, k$. После определения центров кластеров на первом этапе их количество может уменьшиться, поскольку не в каждом тексте можно найти все смысловые аспекты с схемы смысловых аспектов. Итак, для всех последующих итераций $k = k^* \leq N_A$.

Учитывая (5) и (6) целевую функцию

$$F = \sum_{i=1}^{N_{s, d}} \sum_{j=1}^k u_{ij} \text{sim}(s_i, O_j)$$

запишем в виде

$$\begin{aligned} F^* &= \sum_{i=1}^{N_{s, d}} \sum_{j=1}^k u_{ij} \text{sim}(s_i, O_i) - \\ &- \sum_{i=1}^{N_{s, d}-1} \sum_{j=1}^k \sum_{\zeta=i+1}^{N_{s, d}} \sum_{l=1, l \neq j}^k u_{ij} u_{\zeta l} \text{sim}(s_i, s_\zeta). \end{aligned} \quad (8)$$

Максимизация целевой функции (8) обеспечивает максимизацию сходства в пределах кластера и минимизацию сходства междукластерных предложений.

ЗАКЛЮЧЕНИЕ

Был предложен метод реферирования текстов образовательной сферы деятельности на основе схемы содержательных аспектов соответствующего класса документов путем предварительной кластеризации предложений. Применение "min-max" принципа кластеризации обеспечит однородность в пределах содержательных аспектов и обособленность вне содержательных аспектов. Ранжирование предложений документа осуществляется с учетом веса содержательных аспектов.

1. Yeh J. - Y., Ke H. - R., Yang W. - P., Meng I. - H. Text summarization using a trainable summarizer and latent semantic analysis // Information Processing and Management, – 2005. –Vol. 41, № 1. –P. 75–95.
2. Леонов В. П. Реферирование и аннотирование научно-технической литературы / В. П. Леонов; Отв. ред. Б. С. Еленов. – Новосибирск: Наука, 1986. –176 с.
3. Функциональность специализированных информационно-аналитических систем для поддержки информационно-учебной деятельности / В. П. Тарасенко, А. Ю. Михайлюк, М. В. Снежко, Л. М. Бигун // Проблемы информатизации и управления. – Сб. науч. работ. – К.: НАУ, 2009. – № 3 (27). – 123–130 с.