ТАКСОНОМИЗАЦИЯ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ

Виталий Приходнюк

***Аннотация**: В статье описывается подход формирования таксономий на основе семантического анализа текстовых массивов. Представлен алгоритм и описаны основные этапы его работы, определена спецификация входных и выходных данных в виде бестиповых термов ламбда-исчисления. Приведены так же вспомогательные алгоритмы выделения различных типов (в частности, географической) информации. Дается оценка эффективности предложенного алгоритма, полученная с помощью вычислительных экспериментов.*

***Ключевые слова**: таксономия, гиперотношение, структуризация, инженерия знаний*

***ACM Classification Keywords**: I.2 ARTIFICIAL INTELLIGENCE - I.2.4 Knowledge Representation Formalisms and Methods, H. Information Systems*

Введение

Быстрый рост тематических объемов информации, необходимость ее более качественной обработки и усвоения требуют использования методов, средств получения информации и преобразования ее в такую форму, с которой будет удобнее работать на всех этапах решения задач. Главная цель такого преобразования заключается в нахождении документов по нужной тематике, обработке и анализе текстовых (естественно-языковых) документов с помощью определенных инструментов, которые позволяют выявлять свойства описанных объектов и логические закономерности, существующие между ними. По мнению ученых, предложенная должным образом информация позволяет увидеть те дополнительные скрытые закономерности, которые не удается обнаружить другими методами [Величко, 2009; Гаврилова, 2001; Валькман, 2012; Гладун, 1994; Van Rijsbergen, 1979; Helbig, 2006].

Таким образом, актуальной является задача идентификации терминов, которые совместно с контекстами определяют содержание документа и семантические связи между ними. Это дает возможность в дальнейшем сформировать топологическую структуру текста в виде таксономии, отображения и создания новых объектов, связей, увязки новых атрибутов, которые могут быть

использованы при аналитической обработке информации [Стрижак, 2014; Величко, 2015]. Такие топологические структуры могут использоваться при формировании информационной среды корпоративных систем (КС), сетевые инструменты которых обеспечивают поиск, формулировки, формирования, структурирования и представления информации и сообщений, из которых в дальнейшем формируются знания и принимаются соответствующие решения.

Процесс выделения информации

Эффективная обработка неструктурированных (в частности, написанных естественно-языковым текстом) документов может достигаться с помощью их таксономизации [Стрижак, 2014; Шаталкин, 2012] с последующим представлением в виде онтологического графа [Величко, 2009].

На процесс взаимодействия с текстовыми информационными ресурсами, особенно в сетевой среде влияют такие три аспекта, как:

а) синтаксический, который касается формальной правильности сообщений с точки зрения синтаксических правил языка, используемого безотносительно к его содержанию;

б) семантический, который отражает уровень понятийного взаимодействия;

в) прагматический, который определяет операциональные аспекты их использования.

Первичная обработка информационных ресурсов, особенно при их использовании, требует решения целого звена проблем, которые также характеризуют процессы взаимодействия. К этим проблемам специалисты относят следующие: распределенность; гетерогенность; интероперабельность информации только на синтаксическом и структурном уровнях; неполную ответственность за информацию, передаваемую при интеграции; дублирование информации; потерю полноты контроля доступа к информации; технологические трудности, связанные с разнообразием форматов представления данных; содержательность конфликтов между информационными единицами на понятийном уровне; информационная энтропия источника информации. И каждая из этих проблем имеет свои определенные проблемные вопросы с точки зрения технологии ее решения.

В рамках данного процесса необходимо пройти три ключевых *этапа*:

- идентификацию множества терминов концептов X, принадлежащих заданному тексту терминополя;
- идентификацию множества семантических отношений R_{sem} между концептами;

— идентификацию множества атрибутов концептов A (таких, как географическая или темпоральная информация).

Идентификация возможна с помощью последовательного преобразования входного множества лексем естественно-языкового текста L с помощью последовательного применения правил из множества Rul.

На множестве лексем $l \in L$ с помощью оператора «$\prec$» (предшествует) определено линейный порядок, и, таким образом, L является линейно упорядоченной множеством $l_1 \prec l_2 \prec ... \prec l_n$. Также лексемы разбиты на предложения S_i, на множестве которых аналогичным образом задано линейный порядок $S = \{S_1 \prec S_2 \prec ... \prec S_m\}$.

Каждое предложение также представляет собой линейно упорядоченное множество: $S_i = \{l_1^i \prec l_2^i \prec ... \prec l_{n_i}^i\}$. Правила применяются отдельно к каждому из предложений S_i и действуют исключительно в рамках предложения. Сами правила имеют аппликативную форму и могут быть представлены в виде бестиповых выражений [Барендрегт, 1985; Стрижак, 2014]:

$$f_a = (\lambda x.t(x))a = t(a) \quad (1)$$

где:

— λ-теория – ламбда-исчисления; запись λx подразумевает, что это λ-терм;
— x – переменная, принимающая значения на множестве лексем L или концептов X;
— t – выражение, содержащее переменную;
— a – аргумент функции, определяющей возможные значения переменной x;
— f_a – функция, которая может быть применена к аргументу a.

Каждое такое правило задает преобразование одного из видов (2) – (4).

$$L \xrightarrow{Rul} X \quad (2)$$

$$X \xrightarrow{Rul} \langle X, R_{sem} \rangle \quad (3)$$

$$L \xrightarrow{Rul} A \quad (4)$$

Кроме того, возможны другие преобразования:

$$L \xrightarrow{Rul} L \quad (5)$$

$$L \xrightarrow{Rul} L^* \quad (6)$$

где множество L^* – множество конструктов.

Конструкт объединяет в себе несколько лексем, которые в дальнейшем обрабатываются как одна. Конструкты могут иметь такие же характеристики, как и лексемы, и могут быть связаны с другими лексемами или конструктами синтаксическими связями R_{syn} .

Любое правило вида (2), (4), а в некоторых случаях – и вида (5) может быть применено не только к множеству L , но и к множествам L^* или $L \cup L^*$.

Предварительная структуризация

Первым и наиболее очевидным источником структуры текста является его содержание. Оно представляет собой набор предложений $S_{toc} \subset S$, которые определенным образом выделены из основного текста. Чаще всего под содержание отводится несколько страниц в начале или в конце текста. Тогда содержание достаточно легко идентифицировать, задав его пределы, и воспользовавшись гиперотношением множественности порядка S [Клини, 1957; Малишевский, 1998]. Применение гиперотношения S , является необходимым условием и обеспечивает идентификацию конкретных мест в тексте, в которых позиционируются конкретные понятия, и на которые ссылаются элементы содержания. Описываемая процедура разметки текста реализуется на основе следующих двух правил:

$$T = \lambda l_1, l_2, ..., l_n.t \quad (7)$$

$$t \equiv \exists i, \forall j \in [1, n_i], S_i \in S_{toc} \cup l_j^i \in S_i \quad (8)$$

Конструктивность правил (7) и (8) позволяет их применять даже без предварительного использования процедур оригинальной разметки текста, что характеризует процессы формирования лингвистических корпусов [Широков, 2005].

При отсутствии содержания необходимо сформировать предикат q для анализа разметки и заменить условие (8) на (9).

$$t \equiv \exists i, \forall j \in [1, n_i], q\left(l_j^i\right) \quad (9)$$

После применения предиката идентификации выделенные им последовательности лексем формируют множество категорий $\{X_{cat}\}$. Благодаря линейному порядку лексем и предложений можно разбить оригинальный текст на части:

$$L_i^{cat} \equiv \left\{ l \middle| \forall l^{i-1} \in S_{i-1}^{toc}, \forall l^{i+1} \in S_{i+1}^{toc}, l^{i-1} \prec l \prec l^{i+1} \right\} \tag{10}$$

Каждую из множеств L_i^{cat} можно обрабатывать как отдельный текст.

Категории $\{X_{cat}\}$ формируют верхний уровень онтографа: все выделенные из фрагмента текста L_i^{cat} категории являются подкатегориям соответствующей категории X_i^{cat}.

Выделение концептов и связей

Выделение концептов и связей является сложным процессом через большую вариативность языковых конструкций возможных в тексте. Анализатор должен иметь формальное описание таких конструкций, а качество анализа напрямую зависит от полноты этого описания.

Описание предложений в виде правил вида (1). Конкретный вид правил зависит от типа правила и входящего подмножества лексем, для обработки которых предназначено это правило. Составляющими правилами есть предикаты идентификации вида (11) и (12), которые предназначены для обработки отдельной лексемы:

$$c_{a,b} = (\lambda x, y.t(x, y))a, b \equiv \langle a, b \rangle \in LP \tag{11}$$

$$r_{a,b,c} = (\lambda x, y, z.t(x, y, z))a, b, c \equiv \langle a, b, c \rangle \in LS \tag{12}$$

Для каждого предиката определенным образом формируется множество LP или LS. Так LP представляет собой множество лексем и может быть определена двумя способами: простым перечнем допустимых лексем или определением определенного признака, что формирует категорию таких лексем. А LS представляет собой множество пар лексем, связанных определенным видом синтаксической связи. Таким образом, каждый предикат определяется на основе выделения соответствующего ему множества допустимых лексем.

На основе таких предикатов формируется правило вида (13):

$$rul = c_{x_1 p_1} \wedge c_{x_2 p_2} \wedge c_{x_n p_n} \wedge r_{x_1 x_2 k_{12}} \wedge r_{x_2 x_3 k_{23}} \wedge r_{x_{n-1} x_n k_{n-1n}} \quad (13)$$

Применение правила заключается в нахождении упорядоченного множества лексем (14), для которых выполняется условие (15).

$$L_{rul} \subset L, l_1^{rul} \prec l_2^{rul} \prec \ldots \prec l_n^{rul} \quad (14)$$

$$(\lambda x_1, x_2, \ldots, x_n . rul) l_1^{rul}, l_2^{rul}, \ldots, l_n^{rul} \quad (15)$$

Правила вида (2), (4), (5), (6) в дальнейшем выполняют преобразование (16) – (19) соразмерно:

$$\left\{l_1 \ldots l_k, l_1^{rul} \ldots l_n^{rul}, l_{k+n} \ldots l_m\right\} \xrightarrow{rul} \left\{l_1 \ldots l_k, x, l_{k+n} \ldots l_m\right\} \quad (16)$$

$$\left\{l_1 \ldots l_k, l_1^{rul} \ldots l_n^{rul}, l_{k+n} \ldots l_m\right\} \xrightarrow{rul} \left\{l_1 \ldots l_k, \mathrm{a}, l_{k+n} \ldots l_m\right\} \quad (17)$$

$$\left\{l_1 \ldots l_k, l_1^{rul} \ldots l_n^{rul}, l_{k+n} \ldots l_m\right\} \xrightarrow{rul} \left\{l_1 \ldots l_k, l, l_{k+n} \ldots l_m\right\} \quad (18)$$

$$\left\{l_1 \ldots l_k, l_1^{rul} \ldots l_n^{rul}, l_{k+n} \ldots l_m\right\} \xrightarrow{rul} \left\{l_1 \ldots l_k, l^*, l_{k+n} \ldots l_m\right\} \quad (19)$$

Правила формата (3) имеют другую структуру и выполняют преобразование (20):

$$\left\{l_1^{rul}, l_2 \ldots l_{n-1} l_2^{rul}\right\} \xrightarrow{rul} \left\langle \left\{l_1^{rul}, l_2 \ldots l_{n-1} l_2^{rul}\right\}, \left\{\left\langle l_1^{rul}, l_n^{rul}, R_{sem} \right\rangle\right\} \right\rangle \quad (20)$$

Выделение атрибутов

Выделение кандидатов в атрибуты происходит при выделении концептов правила (4). В результате их применения преобразованиями (17) формируется множество A^* . Для формирования множества собственно атрибутов A необходимо осуществить процедуру валидации элементов $a \in A^*$ и отбросить те, которые не пройдут валидацию.

Для каждого из возможных типов атрибутов создается отдельный предикат валидации q , что и определяет, должна ли лексема входить в итоговое множество атрибутов. Предикаты валидации зависят от многих факторов, в частности, типа текста, подмножества языка, обрабатываемого предметной области. Например, для географических координат условием валидности может

быть принадлежность координат определенной рабочей области. Образующий предикат – правило будет выглядеть (21), а сам предикат – вид (22):

$$f_q = x_{\min} < a_x < x_{\max} \wedge y_{\min} < a_y < y_{\max} \tag{21}$$

$$q(a_x, a_y) = \begin{cases} 1, f_q(a_x, a_y) \\ 0, \neg f_q(a_x, a_y) \end{cases} \tag{22}$$

На основе таких предикатов формируются множества атрибутов по типам (23) и обобщающее множество (24):

$$A^i = \left\{ a^i \mid q_i(a^i) \right\} \tag{23}$$

$$A = \bigcup_i A^i \tag{24}$$

Все элементы множества вида (24) могут быть использованы как атрибуты отображения различных массивов обрабатываемых текстов. За счет применения гиперотношения S в виде правил (7), (8), (20) они обеспечивают уникальность представления множественности их смыслов, и могут быть использованы в процедурах поиска и идентификации необходимых текстовых массивов. Также, указанные атрибуты, составляющие множество вида (24), могут быть использованы в процедурах интеграции распределенных текстовых массивов, которые имеют определенную степень смысловой эквивалентности. Более того конструктивность правил (2) – (8) и предиката (22) позволяет формировать процедуры расширения смыслов текстовых массивов при их интеграции, на основе связности их элементов гиперотношением множественности порядка S для всех элементов обрабатываемых текстов.

Построение таксономии текста

Сформированное множество атрибутов вида (24), характеризуется тем, что над всеми его элементами задается гиперотношение множественности порядка S . Тогда на основании

применения предикативного выражения вида (22), в нем всегда можно выделить непустое множество элементов, образующих бинарные пары вида:

$$\lambda((x_i)rul)S\lambda(y_i)rul) \tag{25}$$

каждый терм которого представляет определенную лексему обрабатываемых текстов. Конструктивной особенностью выражения (25) является представимость каждой бинарной пары в виде тематической тавтологии [Стрижак, 2014].

Процесс построения таксономии текста теперь, на основе применения правил (1) – (24), может быть представлен в виде следующей продукции:

$$\lambda((x_i)rul)S\lambda(y_i)rul) \Rightarrow \breve{T} = (\lambda(x.t(x), S, \prec) \tag{26}$$

Правило (26) задает индуктивность процесса формирования упорядоченных множеств концептов вида (24), между элементами которых устанавливаются гиперотношение множественности порядка, и фактически конструируется таксономия. Необходимым условием индуктивности является определение над концептами текстов предикативного выражения (22). Предикативные выражения, формулируются на основе концептов таксономической категории с заданным множественным отношением упорядоченности и принимают только значения истинности. Это позволяет формировать на основе терминов концептов таксономической системы, лингвистические выражения, которые содержательно отражают смысловые состояния текстовых массивов, как пассивной системы тематических знаний.

Так, для множества таксономических категорий = {тип (phylum) подтип (subphylum) класс (classis) подкласс (subclassis) ряд (у растений - порядок) (ordo) подряд (subordo) семья (familia) подсемейство (subfamilia) род (genus) подрод (subgenus) вид (species) подвид (subspecies) разновидность (varietas) форма (forma)}, гиперотношение бинарной множественности порядка, обеспечивает сохранение всех типов взаимодействия между термами, определяющих тематики текстов. Это также позволяет формировать все множества таксономий из концептов

сложившейся онтологической системы. Однако задав над указанными категориями отношение линейной упорядоченности, мы сужаем их перечень, так как ряд категорий, таких, как: класс (classis) форма (forma) вид (species) могут занимать в выражениях (24) и (25), представляющий бинарное отношение - «быть элементом категории», как левую, так и правую часть выражений вида (11), (12) и (26).

В заключение отметим, что согласно [Малишевский, 1998; Стрижак, 2014], бинарные выражения, составляющие правила (24) – (26) обладают свойствами агиперцикличности, иррефлексивности, гипертранзитивности и регулярности:

– агиперцикличность – если для S не существует гиперцикличного множества концептов $X \subseteq U$ такого, когда:

$$\forall x \in X \exists Y \subseteq X : YSx \tag{27}$$

– иррефлексивность:

$$YSx \Rightarrow (Y / \{x\})Sx \tag{28}$$

– гипертранзитивность:

$$YSx, x \in X, XSz \Rightarrow ((Y \cup X) / \{x\})Sz \tag{29}$$

– регулярность:

$$YSx, Y' \supseteq Y \Rightarrow Y'Sx \tag{30}$$

Указанные свойства позволяют в дальнейшем реализовывать процедуры, которые обеспечивают формирование на основе выделенных таксономий, тематических онтологических систем [Гаврилова, 2001; Валькман, 2012; Гладун, 1994; Стрижак, 2014; Guarino, 1994].

Оценка эффективности

Из-за чрезвычайно большой вариативности предикатов и правил оценка эффективности работы алгоритмов вполне может быть выполнена только экспериментальными методами.

Очень чувствительна эффективность к качеству предварительной обработки текста (лексического анализа) и соответствия базы правил языка.

Для оценки качества исчисляются такие параметры:

- *TP (True Positive)* – количество объектов, которые правильно идентифицированы системой;
- *FP (False Positive)* – количество последовательностей лексем, которые не определяют объекты, но были идентифицированы системой;
- *FN (False Negative)* – количество объектов, которые не были идентифицированы системой.

Оценка качества работы алгоритма осуществляется по следующим параметрам [Helbig, 2006]:

- *Precision (точность)* представляет собой отношение корректно идентифицированных объектов количества всех идентифицированных системой объектов;
- *Recall (полнота)* представляет собой отношение количества правильно выделенных идентифицированных системой объектов с количеством всех объектов в тексте;
- *F (мера)* представляет собой интегральный показатель точности и полноты, а также вычисляется как их среднее гармоничное.

Данные параметры определяются по формулам (25) – (27):

$$Precision = \frac{TP}{TP + FP} \tag{31}$$

$$Recall = \frac{TP}{TP + FN} \tag{32}$$

$$F = (\beta^2 + 1)\frac{Precision \times Recall}{\beta^2 Precision + Recall} \tag{33}$$

Поскольку трудно оценить относительную важность точности и полноты в процессе выделения текстов, поэтому есть смысл использовать сбалансированную F-меру (28):

$$F = \frac{Precision \times Recall}{Precision + Recall} \tag{34}$$

Результаты вычисления эффективности работы алгоритма для простого текста, описывающего географическое размещение различных объектов, показано в таблице:

	Точность	Полнота	F-мера
Имена	0,723404	0,85	0,781609
Географическая информация	0,92	0,741935	0,821429
Всего	0,824742	0,784314	0,80402

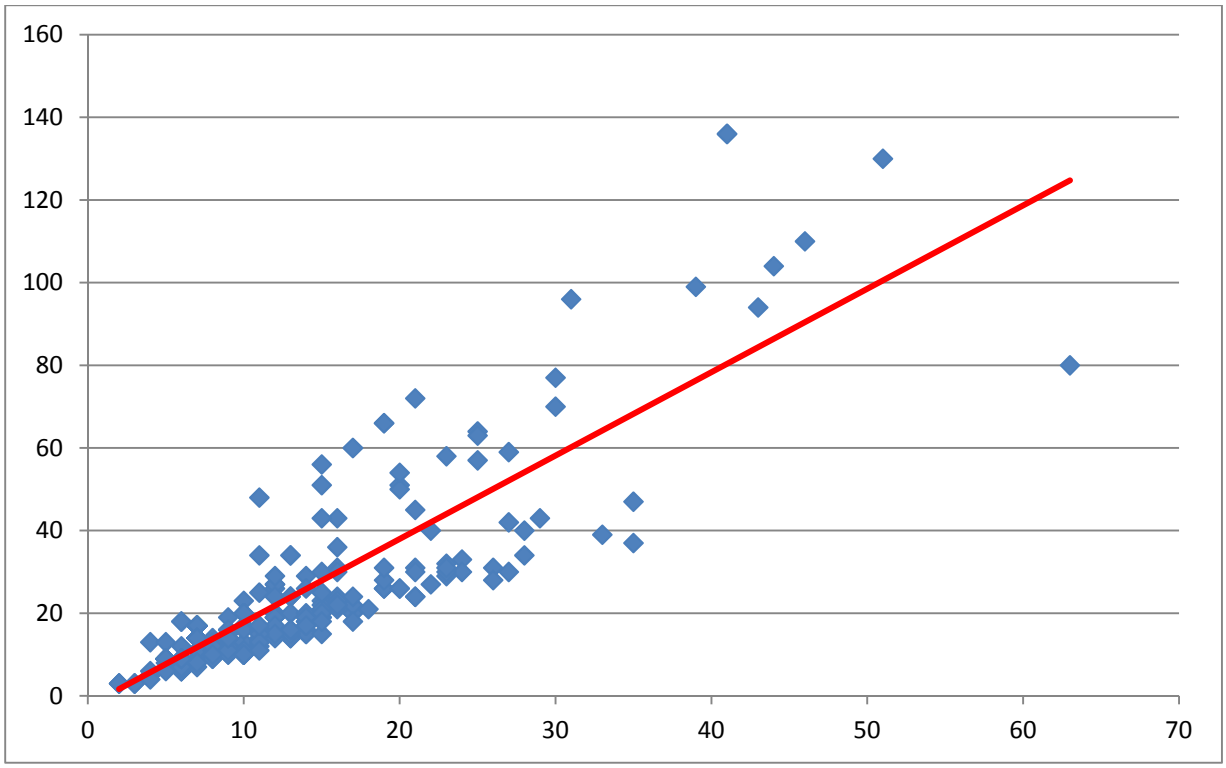
Как видно из таблицы, качество выделения имен (что, по сути, являются концептами) и географической информации (атрибуты концептов) кардинально отличаются.

Географическая информация после идентификации проходит процедуру валидации, которая позволяет добиться чрезвычайно высокой точности. Однако в ходе такой процедуры некоторые выделенные элементы данных отбрасываются, что значительно снижает полноту.

Для имен ситуация прямо противоположная – для них не существует эффективных алгоритмов валидации, поэтому точность их идентификации относительно низкая. Но благодаря тому, что не происходит отвержение элементов данных, повышается их полнота.

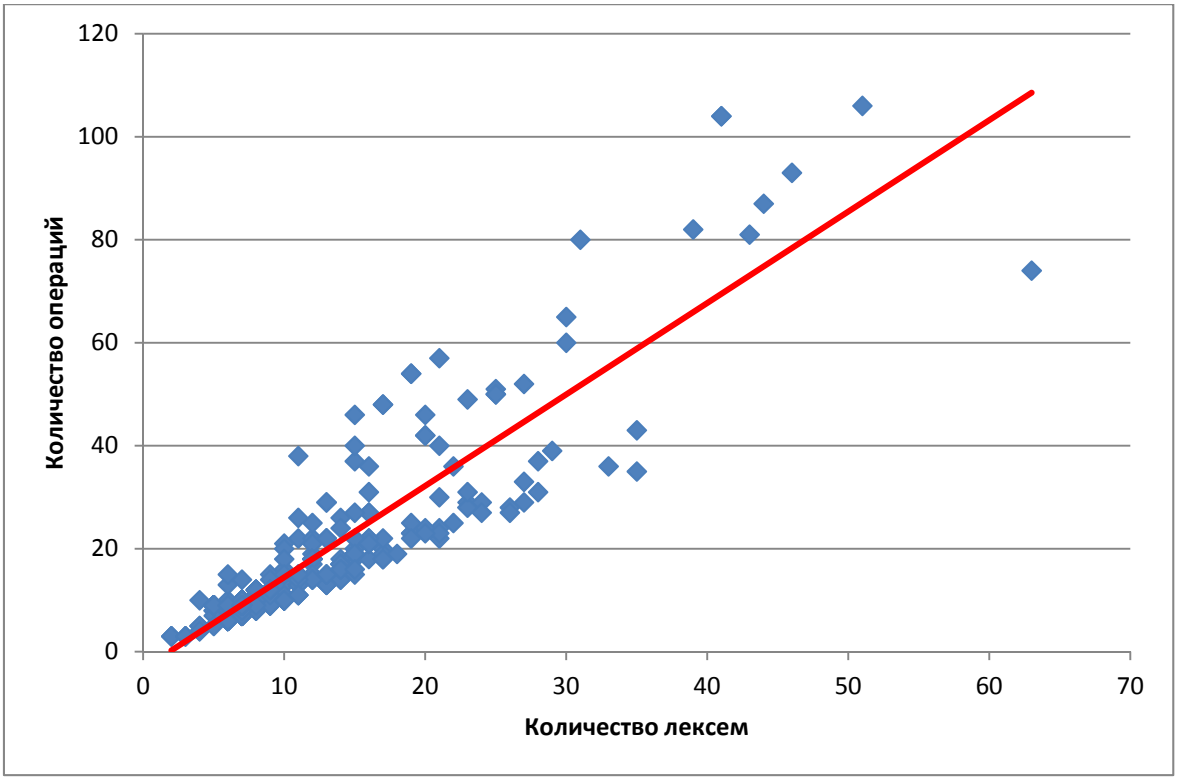
В целом эффективность выделения имен несколько ниже, поскольку в способах написания имен значительно больше вариативности.

Быстродействие работы алгоритма напрямую зависит от количества вызова операций применения предиката к входной лексеме. Поскольку каждое правило работает в рамках одного предложения, то зависимость быстродействия от количества предложений является линейной. Зависимость скорости от длины предложения показано ниже на диаграмме:



Как видно, зависимость количества вызовов операций (и, соответственно, производительности) от длины предложения, близка к линейной. При этом эффективность обработки более распространенных коротких предложений больше, чем менее распространенных длинных.

Также быстродействие зависит от размеров базы правил. Например, если отбросить около половины правил, быстродействие алгоритма несколько повысится, и будет выглядеть следующим образом:



Как видно, основным фактором, влияющим на скорость работы алгоритма, является размер входного текста. Зависимость скорости работы алгоритма от длины текста (в лексемах) показано на графике:

В целом алгоритм имеет достаточно высокое быстродействие и пригоден для использования в процессах поддержки принятия решений, требующих оперативного анализа данных. Алгоритм также пригоден для структуризации больших объемов данных, таких, как книги.

Выводы

Таким образом, применение гиперотношения множественного порядка к множеству понятий, составляющих терминополе текстов, позволяет их представлять как в виде бестиповых термов, описываемых при помощи λ-теории ламбда-исчисления. Бестиповые процедуры обеспечивают процесс идентификации мест позиционирования конкретных понятий текста, выделяют бинарные структуры, которые могут быть также представлены тематическими тавтологиями. За счет этого реализуется структуризация текстовых массивов и формирование их таксономических систем. В общем случае таксономию текстового массива можно определять как архитектуру соответствующего документа. Другими словами таксономия есть обобщающее покрытие тематического многообразия текстового документа.

Конструктивной характеристикой указанного многообразия является ее интегративность и выявление смысловых связностей, представляемых в виде продуцируемых гиперотношением множественности порядка тематических тавтологий и установлением между ними бинарного отношения порядка.

Более того бестиповая множественность порядка определяет достаточно эффективные процедуры разметки и идентификации понятий терминополя текстовых документов и как следствие достаточно эффективное многообразие их таксономических систем .

Bibliography

[Guarino, 1994] Guarino N. The Ontological Level [Текст] / N. Guarino, R. Casati, N. Smith, G. White // Philosophy and the Cognitive Sciences. – Vienna : Holder-Pichler-Tempsky, 1994. – p. 443-456.

[Helbig, 2006] Hermann Helbig: Knowledge Representation and the Semantics of Natural Language [Text]. – Berlin : Springer, 2006. – 651 p.

[Van Rijsbergen, 1979] Van Rijsbergen, C. J. Information Retrieval (2nd ed.). – Butterworth. – 1979. – 208 p.

[Барендрегт, 1985] Барендрегт Х. Ламбда-исчисление. Его синтаксис и семантика: Пер. с англ. – М. : Мир, 1985. – 606 с.

[Валькман, 2012] Валькман Ю. Р. Модельно-параметрическое пространство: теория и применение : [монография] / Ю. Р. Валькман, В. И. Гриценко, А. Ю. Рыхальский. – К. : Наукова думка, 2012. – 192 с.

[Величко, 2009] Величко В. Автоматизированное создание тезауруса терминов предметной области для локальных поисковых систем / В. Величко, П. Волошин, С. Свитла // «Knowledge – Dialogue – Solution» International Book Series «INFORMATION SCIENCE & COMPUTING», Number 15. – FOI ITHEA Sofia, Bulgaria. – 2009. – p. 24–31.

[Величко, 2015] Построение таксономии документов для формирования иерархических слоев в геоинформационных системах [Текст] / Виталий Величко, Виталий Приходнюк, Александр Стрижак, Крассимир Марков, Крассимира Иванова, Стефан Карастанев // International Journal "Information Content and Processing", 2015. – Volume 2. – Number 2. – p.181-199.

[Гаврилова, 2001] Гаврилова Т. А. Базы знаний интеллектуальных систем [Текст] / Т. А. Гаврилова, В. Ф. Хорошевский. – СПб. : Питер, 2001. – 384 с.

[Гладун, 1994] Гладун В. П. Процессы формирования новых знаний [Текст] / В. П. Гладун. – София : СД «Педагог 6», 1994. – 189 с.

[Клини, 1957] Клини, С. К. Введение в метаматематику [Текст] / С. К. Клини. – М. : Иностранная литература, 1957. – 526 с.

[Малишевский, 1998] Малишевский А. В. Качественные модели в теории сложных систем / А. В. Малишевський. – М. : Наука. Физматлит, 1998. – 528 с.

[Стрижак, 2014] Стрижак А.Е. Таксономические характеристики онтологических систем [Текст] / А. Е. Стрижак // Бионика интеллекта, 2014. – № 2(83). – с. 24-29.

[Стрижак, 2014] Стрижак О. Є. Трансдисциплінарна інтеграція інформаційних ресурсів [Текст] : автореф. дис. ... д-ра техн. наук : 05.13.06 / Стрижак Олександр Євгенійович ; Нац. акад. наук України, Ін-т телекомунікацій і глобал. інформ. простору. Київ, 2014. 47 с.

[Шаталкин, 2012] Шаталкин, А.И. Таксономия. Основания, принципы и правила [Текст] / А. И. Шаталкин. – М. : Товарищество научных изданий КМК, 2012. – 600 с.

[Широков, 2005] Широков В. А., Булгаков О. В., Грязнухіна Т. О. та ін. Корпусна лінгвістика [Текст] / В. А. Широков, О. В. Булгаков, Т. О. Грязнухіна та ін. – К.: Довіра, 2005. – 471 с.

Authors' Information

Виталий Приходнюк – аспирант, Институт телекоммуникаций и глобального информационного пространства НАН Украины, Киев-186, 03186, Чоколовский бульвар, 13; e-mail: vitalik1700@yandex.ru

Основные области научных исследований: Data Mining, геоинформационные системы, онтологический инжиниринг

Taxonomyzation of Natural Language Texts

Vitaly Prihodnyuk

Abstract*: An approach for forming taxonomies based on semantic analysis of text arrays is presented in this paper. The algorithm and the main stages of its work are described. The specification of input and output data is defined as lambda calculus terms without types. Auxiliary algorithms for detecting different types (in particular, geographic) information are outlined. The evaluation of the effectiveness of the proposed algorithm, obtained by computational experiments, is given.*

Keywords*: Taxonomy, Hyper-relation, Structuring, Knowledge Engineering*

ACM Classification Keywords*: I.2 ARTIFICIAL INTELLIGENCE - I.2.4 Knowledge Representation Formalisms and Methods, H. Information Systems*