УДК 519.688

ВЫДЕЛЕНИЕ ТЕРМИНОВ И ИХ СВЯЗЕЙ ДЛЯ ПРЕДМЕТНОГО УКАЗАТЕЛЯ НАУЧНОГО ТЕКСТА

Е.И. Большакова (*bolsh@cs.msu.ru*)
МГУ имени М.В. Ломоносова, НИУ ВШЭ, Москва

К.М. Иванов (*ivanov.kir.m@yandex.ru*)
МГУ имени М.В. Ломоносова, Москва

Предметный указатель – список значимых терминов текстового документа с указанием страниц, на которых они употребляются. В работе описываются методы на основе лексико-синтаксических шаблонов и правил, разработанные для автоматического извлечения и отбора терминов в предметный указатель заданного научного текста, а также для выявления их подчинительных связей.

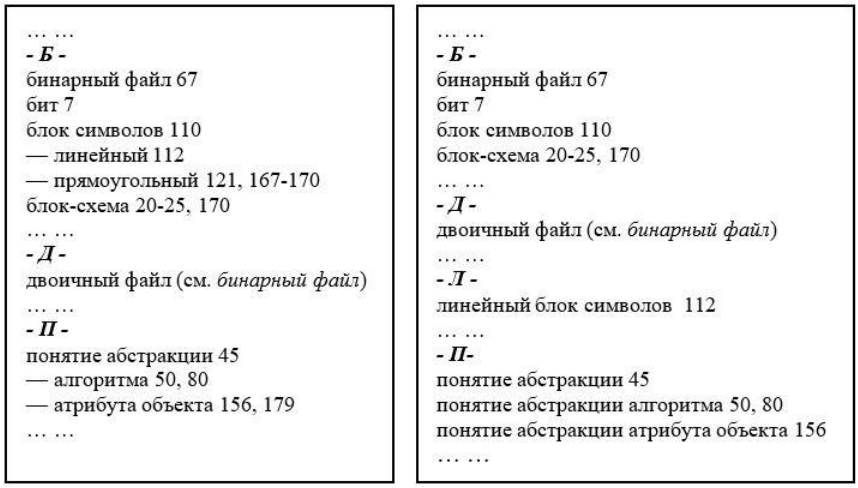
Ключевые слова: извлечение терминов, предметный указатель, научные тексты, лексико-синтаксические шаблоны и правила

Введение

Предметные указатели (back-of-the-book indexes) создаются для текстовых документов большого и среднего объема, таких как книги, руководства, учебные пособия и др., особенно в узкоспециализированных областях. Предметный указатель (ПУ) содержит значимые термины текста с указанием номеров страниц, где они употребляются, облегчая тем самым поиск нужной информации. Типичные фрагменты ПУ показаны на Рис. 1: соответственно иерархического указателя (слева) и плоского (справа). Иерархический указатель включает кроме самих терминов их подчинительные (иерархические) связи (например: *блок символов – линейный блок символов*), и его можно рассматривать как квазионтологию понятий, рассмотренных в документе.

Из-за трудоемкости создания ПУ они часто отсутствуют в книгах, особенно это касается текстов из быстро развивающихся научно-технических областей, что делает задачу автоматизированного их построения актуальной. В тоже время проблемы автоматизации построения предметного указателя для заданного текстового документа слабо изучены. Центральной проблемой является извлечение однословных и многословных терминов из заданного текста на основе лингвистических

и статистических критериев и последующий отбор наиболее подходящих для ПУ. Для построения иерархического предметного указателя также требуется выявление смысловых связей уже извлеченных терминов: подчинительных и синонимичных, последние представляются в указателях перекрестными ссылками (на Рис. 1 это связь *двоичный файл – бинарный файл*).



… …
- Б -
бинарный файл 67
бит 7
блок символов 110
— линейный 112
— прямоугольный 121, 167-170
блок-схема 20-25, 170
… …
- Д -
двоичный файл (см. *бинарный файл*)
… …
- П -
понятие абстракции 45
— алгоритма 50, 80
— атрибута объекта 156, 179
… …

… …
- Б -
бинарный файл 67
бит 7
блок символов 110
блок-схема 20-25, 170
… …
- Д -
двоичный файл (см. *бинарный файл*)
… …
- Л -
линейный блок символов 112
… …
- П-
понятие абстракции 45
понятие абстракции алгоритма 50, 80
понятие абстракции атрибута объекта 156
… …

Рис. 1. Фрагменты предметного указателя

Широко применяемые для извлечения терминов методы и технологии [Korkontzelos et al., 2012] опираются в основном на статистику терминоупотреблений и ориентированы на обработку коллекций специализированных текстов, с целью создания терминологических словарей, тезаурусов и онтологий соответствующих предметных областей. Однако для извлечения терминов из отдельно взятых текстов они недостаточно эффективны, и решение этой задачи применительно к созданию ПУ рассматривалось в очень небольшом числе работ.

Методы извлечения терминов, разработанные в [Csomai et al., 2007, 2008] для построения ПУ англоязычных документов, используют структурные особенности терминов и различные статистические меры, основанные на частоте встречаемости слов в тексте, но даже при применении машинного обучения они достигают лишь 27-28% полноты и точности. Для схожей задачи извлечения ключевых слов результаты, приведенные в статье [Hasan et al., 2014], составляют 35% точности, 66% полноты и 45,7% F-меры (комбинированного показателя полноты и точности). Примерно такие же оценки дает метод извлечения терминов

для ПУ, описанный в работе [Wu et al, 2013] и опирающийся на комбинацию нескольких статистических мер значимости терминов.

Предложенный в [Arora et al., 2016] метод извлечения терминов для построения глоссариев документов, с применением кластеризации извлеченных терминов, показывает в экспериментах 35-67% F-меры (точность – 21-51%, полнота – около 90%).

В работе [Zargayouna et al., 2006] кратко охарактеризована система InDoc, автоматизирующая построение ПУ, и применяемый метод извлечения терминов, основанный на лингвистических правилах. Правила определяют различные грамматические структуры терминов и их возможные текстовые варианты. Однако оценка эффективности метода в статье не представлена, как и в статье [Da Sylva et al, 2013], также посвященной автоматизации построения предметных указателей.

В настоящей работе, продолжающей исследования [Bolshakova et al., 2015], рассматривается задача извлечения терминов и их связей для построения предметного указателя научно-технического текста, что позволяет учесть особенности употребления терминов в таких текстах и достичь достаточно высокой эффективности их извлечения (в среднем до 70-79% точности, полноты и F-меры). Разработанные методы основаны на лексико-синтаксических шаблонах и правилах, описывающих структуру терминов, а также типичные конструкции их использования в научно-технических текстах на русском языке. Для русскоязычных текстов эта задача решена впервые, а новизна предложенной процедуры отбора терминов в ПУ состоит в учете разных факторов значимости терминов.

Шаблоны и правила формализованы на языке LSPL [LSPL], а методы реализованы с помощью инструментов поддержки этого языка и встроены в исследовательский прототип системы автоматизированного построения ПУ, позволяющей человеку-эксперту редактировать результаты.

1. Шаблоны для извлечения терминов из текста

Для выделения терминов из исходного текста были созданы три набора LSPL-шаблонов и правил.

Первый набор из 12 правил определяет извлечение терминов по их грамматической структуре. Правила фиксируют часть речи составных слов термина и их грамматические характеристики (падеж, число и т. п.). Например, шаблон *A N1 N2 <c = gen>* включает: *N1* – существительное, *A* – прилагательное, *N2* – существительное в родительном падеже (*прямоугольный блок символов*).

53 правила второго набора охватывают большинство типичных для русского языка фраз-определенийтерминов, встречающихся в научных текстах, например: *Плоской триангуляцией мы называем связный плоский граф, каждая грань которого*.... Правила включают как отдельные

лексические единицы таких фраз (глаголы *называть*, *определять* и т.д.), так и вспомогательные шаблоны грамматической структуры терминов (из первого набора правил). В частности, для вышеприведенного примера определения, термин *плоская триангуляция* извлекается следующим правилом: *Term <c = ins> "мы" "называем" => #Term*

где исходный термин *Term* должен быть в творительном падеже (*c = ins*), но извлекается (=>) он в нормальной форме, т.е. как лемма *(#Term*).

Третий набор правил состоит из 25 правил, учитывающих типичные конструкции введения терминологических синонимов и сокращений в русскоязычных научных текстах, например: …*информационная система, или просто ИС*. Правила данной группы распознают и извлекают пары терминологических синонимов, опираясь на знаки пунктуации (запятые) и лексические маркеры (в частности, слова *или просто*). Приведем пример такого правила (слово *просто* задается как опционально):

Term1 "," "или" ["просто"] Term2 = > #Term "-» #Term2

В результате независимого применения рассмотренных наборов правил из текста извлекаются три множества *терминов-кандидатов*, причем эти множества пересекаются, в частности, есть термины, извлеченные как шаблонами грамматической структуры, так и шаблонами конструкций определений терминов.

Для каждого набора (вида) правил была экспериментально определена точность извлечения терминов, для этого использовались два учебно-научных пособия по языкам программирования, включающих авторские предметные указатели. Для первого набора правил эксперименты ожидаемо показали высокую полноту, но низкую точность извлечения (около 8-10%). Правила второго набора, напротив, демонстрируют высокую точность (90-95%), в основном за счет задаваемых в них лексических маркеров. Аналогично, третий набор правил показывает довольно хорошую точность в 63-67%. Точность правил второго набора учитывалась при объединении извлеченных множеств терминов-кандидатов, для отбора наиболее значимых в предметный указатель.

2. Процедура отбора терминов

По результатам нескольких экспериментов с множествами извлеченных терминов, была разработана эвристическая процедура, выявляющая термины предметного указателя.

В начале работы процедура осуществляет фильтрацию терминов-кандидатов, для чего используются два предварительно сформированных списка стоп-слов. Первый список содержит слова, которые сами не могут быть терминами (например: *метод, конец* и т.п.), второй же список включает слова, в основном прилагательные, которые не могут быть

частью терминов (*данный, известный* и т.п.). Из всех трех множеств исключаются слова и сочетания, которые: а) встречаются в первом списке стоп-слов; б) содержат слова из второго списка; с) состоят только из слов первого списка. Тем самым отбрасывается значительное число выражений общенаучной лексики (например: *проблема*, *простая задача* и т.п.).

Затем процедура итеративно формирует набор терминов предметного указателя из элементов уже отфильтрованных множеств, при этом учитываются следующие факторы значимости отбираемых терминов:

- Точность шаблона определения термина, примененного для его извлечения: поскольку новые термины (определяемые в тексте) очевидно относятся к значимым, они подлежат обязательному включению в ПУ, при условии, что соответствующий шаблон принадлежит группе наиболее точных шаблонов;
- Использование термина в заголовке/подзаголовке текстового документа также указывает на его значимость;
- Частота появления термина в тексте: согласно закону Ципфа, наиболее значимыми являются единицы со средней частотой и поэтому они в первую очередь выбираются для предметного указателя;
- Лексическое сходство терминов (мы определяем его как наличие у них общих слов, например: *однородный многочлен* и *многочлен второй степени*): термин-кандидат добавляется в указатель, если он имеет лексическое сходство с любым элементом, уже включенным в ПУ.

Для оценки разработанной процедуры были взяты 5 учебно-научных пособий среднего размера (каждое около 20 тыс. слов), посвящённых языкам (ЯП), системам программирования (СП), а также эвристическим методам поиска (ЭП) в искусственном интеллекте. Включенные в них авторские предметные указатели рассматривались как эталонные множества терминов, и с их помощью оценивались полнота, точность и F-мера применяемого метода. Результаты представлены в Таблице 1.

Табл. 1. Оценка метода отбора терминов для предметного указателя

Текст	Слов в тексте	Извлечено терминов для ПУ		P	R	F
		Число	Примеры			
ЯП (Лисп)	21060	140	*рекурсивная функция*	0.74	**0.84**	**0.79**
ЯП (Рефал)	29301	208	*функциональный терм*	0.56	0.82	0.67
ЯП(Пролог)	21376	77	*предикаты ввода*	**0.77**	0.72	0.78
ЭП	19471	98	*альфа-бета процедура*	0.71	0.74	0.73
СП	11699	67	*языковое окружение*	0.70	0.81	0.75
Среднее	20581	118		**0.70**	**0.79**	**0.74**

Приведенные оценки показывают хорошую эффективность отбора терминов в предметный указатель: для разных текстов полнота составила от 0,72 до 0,84, тогда как точность варьируется от 0,56 до 0,77, что значительно превышает оценки методов отбора (преимущественно статистических и с использованием машинного обучения), описанных в [Csomai et al., 2007, 2008], а также превышает оценки результативности известных открытых систем автоматического извлечения терминов, представленные в работе [Arora et al., 2016]: 20-47% F-меры.

Проведенный анализ случаев неполного или неточного отбора терминов в ПУ показал, что одной из главных причин являются ограничения применяемых лингвистических правил и лексико-синтаксических шаблонов. В частности, некоторые термины не извлекаются из-за их сложной или необычной грамматической структуры, не представленной в текущем наборе шаблонов (к примеру, термин *поиск в ширину* с грамматическим образцом *N + Prep + N*).

Обнаружилось также, что некоторые термины, отобранные для ПУ нашим методом и отсутствующие в эталонном (авторском) предметном указателе (например, *логическое программирование* из текста по языку программирования Пролог), являются тем не менее релевантными для предметного указателя. Они могли быть опущены автором ПУ по разным причинам, так как стандарты на размер и состав предметных указателей отсутствуют. Ясно, что при отборе терминов в предметный указатель полнота извлечения терминов оказывается важнее, чем точность (при условии, что количество извлеченных терминов не слишком велико), поскольку пользователю автоматизированной системы построения ПУ легче отбросить уже выделенные термины, чем искать в тексте и добавлять новые.

3. Выявление связей терминов

Для формирования иерархической структуры предметного указателя текста должны быть выявлены подчинительные связи между парами отобранных в ПУ терминов: внутри указателя один элемент выявленной пары будет представлять *заголовок*, а другой – *подзаголовок* (к примеру, на Рис. 1: *блок символов – линейный*). Подчинительная связь в большинстве случаев оказывается родовидовой связью соответствующих понятий, т.е. члены пары – это гипоним и гипероним, например: *понятие абстракции алгоритма* и *понятие абстракции*.

Отметим, что выявление иерархических связей терминов особо полезно для пользователей предметного указателя, когда среди отобранных единиц много лексически схожих, а значит, и понятийно близких терминов.

Кроме подчинительных связей терминов требуется распознавание синонимических смысловых связей для определения перекрестных ссылок в указателе (см. *двоичный файл – бинарный файл* на Рис.1). В нашем методе пары синонимов распознаются на этапе извлечения терминов, поэтому ниже мы описываем только выявление подчинительных связей на основе информации о структуре многословных терминов.

Известно, что термины-гипонимы часто получаются из терминов-гиперонимов путем дополнения последних определяющими словами [Скороходько, 2001], например: *левая свертка – свертка*, *протокол передачи – протокол передачи сообщения*. Как и в работе [Zargayouna et al., 2006] – единственной известной нам работе, описывающей выявление подчинительной связи терминов для ПУ – мы определяем потенциальные гиперонимы на основе грамматической структуры терминов, используя уже созданный набор лексико-синтаксических шаблонов. Примеры соответствующих грамматических образцов потенциальных гиперонимов для трех разных грамматических структур многословных терминов представлены в Таблице 2.

Табл. 2. Грамматические образцы гипонимов и гиперонимов

Граммат. структура терминов	Пример термина	Потенциальные гиперонимы	
N1 N2<c=gen>	*База знаний*	*N1*	*База*
		N2	*Знания*
A N	*Абсолютная адресация*	*N*	*Адресация*
A1 A2 N	*Абстрактная семантическая сеть*	*N*	*Сеть*
		A1 N	*Абстрактная сеть*
		A2 N	*Семантическая сеть*

Кроме указанной информации о грамматических образцах гиперонимов, представленной в виде шаблонов, используется также простая статистика терминоупотреблений в обрабатываемом документе.

Для выявления подчинительных связей каждого термина T, отобранного для предметного указателя, выполняется следующая процедура. Сначала из состава термина T извлекаются все слова и словосочетания $\{T_p\}$, которые потенциально могут стать для него гиперонимами. Затем вычисляется частота встречаемости в тексте самого T, а также каждого элемента из $\{T_p\}$. Наконец, применяется правило: в качестве гиперонима (заголовка ПУ) выбирается элемент из $\{T_p\}$ с наивысшей частотой встречаемости в тексте – при условии, что его

частота больше частоты встречаемости самого T (если таких элементов несколько, то выбирается первый по алфавиту). В случае, когда частоты всех элементов из $\{T_p\}$ не превышают частоту термина T, то он сам образует самостоятельную единицу указателя (без гипонимов). Указанное правило гарантирует самостоятельное и частотное употребление в тексте термина-гиперонима, что является основанием введения в предметный указатель соответствующего заголовка.

Например, для термина *абстрактная семантическая сеть*, потенциальными гиперонимами являются: *сеть*, *абстрактная сеть*, *семантическая сеть*. Если термин *семантическая сеть* имеет самую высокую частоту, он становится заголовком, а термин *абстрактная семантическая сеть* будет считаться его гипонимом (подзаголовком).

После выявления подчинительных связей терминов формируется иерархическая структура ПУ: из каждого термина-гипонима удаляется его составная часть, идентичная гипериниму, а остальная часть располагается на втором уровне иерархии. Для рассмотренного примера результирующая структура имеет вид: *семантическая сеть*
– *абстрактная*

Поскольку стандарты на построение заголовков в иерархическом предметном указателе отсутствуют, описанный метод выявления подчинительных связей терминов экспериментально не оценивался.

Заключение

В данной работе описан основанный на лингвистических шаблонах и правилах метод извлечения терминов из научно-технических текстов и метод выявления их смысловых связей (подчинения и синонимии), разработанные для автоматизированного построения предметных указателей. Экспериментальная оценка метода извлечения и отбора терминов показала его достаточную эффективность для применения в рамках системы поддержки построения предметных указателей: в среднем 70% точности, 79% полноты и 74% F-меры.

Описанные методы встроены в прототип системы поддержки построения предметного указателя, с открытым программным кодом[1].

Перспективы совершенствования методов связаны с расширением используемого набора шаблонов и правил, а также с уточнением используемых в методах эвристических процедур.

Список литературы

[Скороходько, 2001] Скороходько Э.Ф. Термины, выражающие новые знания в структуре научных текстов // НТИ, Сер. 2., 2001, №4.

[1] https://github.com/ivanov-kir-m/SISTool

[Arora et al, 2016] Arora C., Sabetzadeh M., Briand L., Zimmer F. Automated Extraction and Clustering of Requirements Glossary Terms // IEEE Transactions on Software Engineering, Vol.43, Issue 10, pp. 918-945.

[Bolshakova et al, **2015]** Bolshakova E.I., Efremova N.E. A Heuristics Strategy for Extracting Terms from Scientific Texts // Fourth Int. Conference AIST, CCIS, Vol. 542. Springer Berlin Heidelberg, p. 285-295.

[Csomai et al, 2007] Csomai A., Mihalcea R. Investigations in Unsupervised Back-of-the-Book Indexing // Proc. of the Florida Artificial Intelligence Research Society Conference, pp. 211-216.

[**Csomai et al, 2008**] Csomai A., Mihalcea R. Linguistically Motivated Features for Enhanced Back-of-the Book Indexing // Proceedings Annual Conf. of the ACL, ACL/HLT, Vol. 8, pp. 932–940.

[Da Sylva et al, 2013] Da Sylva L. Integrating Knowledge from Different Sources for Automatic Back-of-the-Book Indexing // Proceedings of the Annual Conference of CAIS/Actes du congrès annuel de l'ACSI.

[Hasan et al, 2014] Hasan K.S., Ng V. Automatic keyphrase extraction: a survey of the state of the art // Proceedings of the 52th Annual Meeting of the ACL, pp. 1262–1273.

[Korkontzelos et al, 2012] Korkontzelos I., Ananiadou S. Term Extraction // Oxford Handbook of Computational Linguistics (2nd Ed.), Oxford University Press, Oxford.

[LSPL] Lexico-Syntactic PatternLanguage – http://lspl.ru/

[Wu et al, 2013] Wu Z. et al. Can Back-of-the-Book Indexes be Automatically Created? // Proceedings of the 22nd ACM Int. Conference on Information & Knowledge Management. pp.1745-1750.

[Zargayouna et al, 2006] Zargayouna H., El Mekki T., Audibert L., Nazarenko A. IndDoc: an Aid for the Back-of-the-Book Indexer // The Indexer, 25(2), pp. 122–125.

EXTRACTION OF TERMS AND THEIR LINKS FOR SUBJECT INDEX OF SCIENTIFIC TEXT

E.I. Bolshakova (*bolsh@cs.msu.ru*)

K.M. Ivanov (*ivanov.kir.m@yandex.ru*)

Moscow State Lomonosov University, Moscow

Subject (or back-of-the-book) index is a list of significant terms from a text document, indicating pages on which they are used. The work describes methods based on lexico-syntactic patterns and rules and developed for extraction of terms and revealing their subordinate links for the subject index of a given scientific text.

Keywords: term extraction, back-of-the-book index, subject indexing, scientific texts, lexico-syntactic patterns and rules