РОЛЬ МАШИННОГО ОБУЧЕНИЯ В ОБРАБОТКЕ ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ТЕКСТОВ

THE ROLE OF MACHINE LEARNING IN PROCESSING NATURAL LANGUAGE TEXTS

Найденова К.А. mailto: naidenova@mail.spbnit.ru

Военно-медицинская академия, Санкт-Петербург

В статье рассматриваются проблемы автоматизированного извлечения лингвистических знаний из естественно-языковых текстов при взаимодействии двух процессов: процесса структурного анализа текста и процесса машинного обучения с учителем, направляющим конструирование необходимых для понимания текста концептуальных объектов.

Извлечение знаний из ЕЯ текстов

Понимание ЕЯ текстов основано на использовании, как лингвистических знаний, так и знаний о мире. Есть два пути конструирования необходимых знаний при автоматизации процессов понимания ЕЯ текстов на компьютере: а) знания представляются экспертами в готовом виде для применения с помощью заранее выбранных средств представления знаний, б) знания извлекаются из ЕЯ текстов с помощью механизмов анализа текстов, реализованных программно, и методов машинного обучения.

Второй путь предполагает, что обучение происходит при взаимодействии двух процессов – «внутреннего» процесса восприятия и структурного анализа ЕЯ текста и процесса обучения с помощью учителя, направляющего восприятие, анализ текста и конструирование необходимых для понимания ЕЯ текстов знаний.

Мы будем исходить из того, что все элементы ЕЯ суть конструктивные объекты (КО). Так, слова конструируются из букв алфавита, предложение состоит из слов, текст из предложений. Кроме того, каждый КО имеет сложную структуру. Например, слово в своем составе имеет основу, префиксы, суффиксы и окончание (возможно пустые). Слова делятся на слоги. Предложение может быть простым и сложным, и т.п. Механизмы «восприятия» и анализа конструктивных ЕЯ объектов должны позволять выделение любых структурных частей этих объектов.

Механизмы обучения должны позволять:

а) выделение концептов - групп объектов (структурных составляющих) и их называние учителем;

б) формирование характерных признаков для выделенных концептов, позволяющих отличать их от других концептов, и их называние учителем;

в) формирование классификаций концептов на основе связей между ними – ассоциативных, импликативных, функциональных, структурных и т.п.

Называние объектов и признаков учителем очень важная составляющая обучения. Если группа объектов (признак) с точки зрения учителя имеет имя, значит, она имеет смысл как концепт. Если нельзя дать имя некоторой группе объектов, значит, это не имеет смысла для учителя, и признаки такой группы объектов можно сохранить как запрещающие правила при формировании концептуальных знаний. Учитель может давать примеры или контр примеры концептов и, тем самым, ускорять процесс обучения компьютера.

Определение КО базируется на задании конструктивного процесса, его порождающего. Например, слово некоторого языка как конструктивный объект, определяется в [1] следующим образом. Пусть А - какой-либо алфавит. Словами в алфавите А называются КО, получающиеся в результате развёртывания конструктивных процессов, ведущихся на основе следующих правил: 1) пустое слово есть слово в алфавите А; 2) если Р слово в алфавите А, то Рξ есть КО, где ξ - любая буква алфавита А.

Правила порождения КО есть также и правила распознавания этих объектов.

Итак, чтобы автоматизированное извлечение концептуальных знаний было возможным, необходимо реализовать два универсальных механизма: 1) выделение структурных элементов всех ЕЯ КО и 2) механизм формирования концептов и структур концептов. Второй механизм есть обучение, в основе которого лежат универсальные механизмы классификации.

Оба механизма известны с математической точки зрения. Первый из них есть «Нормальный алгорифм Маркова» [1]. Второй механизм есть индуктивный вывод закономерностей из данных (Machine Learning) на основе математического аппарата теории алгебраических решеток. Автору неизвестны проекты, в которых применялись бы оба эти взаимосвязанные механизмы для целей автоматизированного извлечения знаний из ЕЯ текстов. Хотя методы машинного обучения на основе аппарата теории решеток уже широко применяются как нашими учеными [2 - 4], так и зарубежными (извлечение ассоциативных, импликативных и функциональных зависимостей из данных [5, 6], концептуальный анализ Р. Вилле [7] мн. др).

В данной статье даются два примера извлечения лингвистических знаний с применением обучения. В первом примере, который более подробно описан в [8], порождаются правила образования наречий образа действия от прилагательных во французском языке. Во втором примере по заданному тексту порождаются основные темы, затрагиваемые в тексте, и по каждой теме формируется краткое резюме или контекст темы.

Алгорифм Маркова и выделение структурных признаков слов

Идеальной моделью анализа КО и построения правил образования таких КО языка, как слова, является нормальный алгорифм Маркова. С его помощью возможно выделять все начала и концы слова, по заданному началу слова определять его концевое дополнение, по заданному концу слова определять дополняющее его начало, определять вхождение одного слова в другое и выявлять структуру этого вхождения.

Приведем основные обозначения и операции алгорифма Маркова [1].

Слова мы будем обозначать вербальными переменными P,Q,R,S, буквы - литеральными переменными ξ, ψ, ζ.

Алфавит ::= буквы;
Слово ::= набор букв;
Пустое слово = пустой набор ::= Λ;
Слово ::= P | Pξ, где P - слово, ξ - буква; $\Lambda\xi=\xi$.
Графическое равенство слов =;
Графическое неравенство слов $\neq$;
Соединение слов [P,Q]; [P, empty]::= P; [P,Qξ]::=[P,Q]ξ .
Операция обращения [Λ! = Λ;
[Pξ ! = ξ[P ;
[ξ! = ξ ;
[PQ!= [Q![P!;
[[P!!=P;
Q =ξP, ξ -первая буква, Q=Pξ, ξ -последняя буква ;
Q =PX, P- начало слова, Q = XP, P - конец слова ;
XY = Z, следовательно, X - начало Z,
(X $\leftarrow$Z) = Y, Y - концевое дополнение X в Z;
YX = Z, следовательно, X - конец Z,
(X$\rightarrow$Z) = Y, Y - начальное дополнение X в Z;
[Xd - длина слова X;
[P(Б) - проекция слова P на алфавит Б.

Просматривая списки начал пары слов, мы сможем выяснить, имеют ли эти слова непустое общее начало. Легко определяется наибольшее общее начало (НОН) (аналогично, наибольший общий конец (НОК)) пары слов. Для нахождения вхождения одного слова в другое и структуры этого вхождения, если оно имеет место, необходимы следующие определения.

Мы будем говорить, что слова P,Q взаимно просты слева, если не существует непустого слова являющегося началом как P, так и Q.

Каковы бы ни были слова P и Q, могут быть построены слова R, S и T такие, что (**1**) P=RS, (**2**) Q = RT и что S взаимно просто слева с T.

Каковы бы ни были слова P и Q, существует единственная тройка слов R, S и T, удовлетворяющая условиям (**1**) и (**2**) и такая, что S взаимно просто слева с T.

Каковы бы ни были слова P и Q, слово R в единственной тройке слов R, S,T, удовлетворяющей условиям (**1**) и (**2**) и такой, что S взаимно просто слева с T, мы будем называть **наибольшим общим началом (НОН)** слов P и Q.

Каковы бы ни были слова P и Q, существует единственное НОН этих слов.

Всякое общее начало слов P и Q есть начало их НОН. Всякое начало НОН двух слов есть их общее начало.

Аналогично определяется **наибольший общий конец (НОК) двух слов.**

Всякое непустое слово допускает единственное представление в виде Pξ и единственное представление в виде ξP. Букву ξ в единственном представлении непустого слова в виде Pξ мы будем называть последней буквой слова. Букву ξ в единственном представлении непустого слова в виде ξP мы будем называть первой буквой слова.

P входит в Q (Q содержит P), если существует пара слов R, S, такая что Q = RPS.

Если существует пара слов R, S, такая что Q = RPS, то существует U такое, что P есть начало U и U есть конец Q.

Каково бы ни было U, если P - начало U, а U - конец Q, то P входит в Q.

Если существует пара слов R, S, такая что Q = RPS, то существует T такое, что P есть конец T и T есть начало Q.

Каково бы ни было T, если P - конец T, а T - начало Q, то P входит в Q.

Составляем список всех концов слова Q и для каждого конца слова составляем список всех его начал. Объединение всех последних списков даёт список всех слов, входящих в Q.

Определение вхождения слова P в слово Q, Q = R*P*S: R - левое крыло, P - основа , S - правое крыло этого вхождения.

Вхождение с пустым левым крылом мы будем называть начальным вхождением. Вхождение с пустым правым крылом мы будем называть концевым вхождением.

Начальное вхождение: *P* (P $\leftarrow$ Q);

Концевое вхождение: (P $\rightarrow$ Q) *P*.

При формировании правил образования французских наречий из французских прилагательных были выявлены следующие структуры вхождений:

Наибольшая общая часть пары слов (назовём её основой) является их НОН:

А. Слово 1 = основа + конец 1;
Слово 2 = основа + конец 2.

Б. Слово 1 = основа;
Слово 2 = основа + конец 2.

В. Слово 1 = основа;
Слово 2 = основа;
Слово 1 = слово 2.

Наибольшая общая часть пары слов (назовём её окончанием) является их НОК:

Г. Слово 1 = начало1 + окончание;
Слово 2 = начало2 + окончание.

Д. Слово 1 = окончание;
Слово 2 = начало2 + окончание.

Е. Слово 1 = окончание;
Слово 2 = окончание;

. Слово 1 = слово 2.

Формирование правил образования наречий от прилагательных с помощью обучения по примерам

Начальные условия для реконструкции правил образования наречий от прилагательных были следующими: фиксирован заданный алфавит, дано разбиение алфавита на гласные и согласные, даны (с помощью списка примеров) три непустые класса слов: наречия образа действия (Н), прилагательные мужского и женского рода (ПМР и ПЖР), от которых эти наречия образованы.

Сравнивались пары слов из всех пар классов: ПМР - ПМР, ПЖР - ПЖР, ПМР - Н, ПЖР - Н, Н - Н. С помощью алгорифма Маркова определялись структурные признаки слов. Выделялись структурные признаки, которые охватывали наибольшие группы слов. Если структурный признак получал имя со стороны учителя, то этот признак запоминался. Например, было выделено окончание ***ment*** у наречий образа действий.

Затем анализировались структуры вхождений для пар слов из следующих пар классов ПМР – ПЖР, ПМР – Н, ПЖР – Н.

Наибольшее количество начальных вхождений в наречия имеют ПЖР, при этом структура наречия имеет вид:

ПЖР + ment.

Выделен класс ПЖР, имеющих структуру:

ПМР + окончание "e".

Причем, в этой структуре ПМР оканчивается на гласную.

Получена также следующая структура наречия:

ПМР - nt + mment.

Первая структура соответствует следующим правилам образования наречий:

1) от прилагательного единственного числа, имеющего одинаковую фонетическую и орфографическую форму мужского и женского рода, оканчивающуюся на немое "e": **large - large - largement** (широко), **intense - intense - intensément** (интенсивно); **commode - commode - commodément** (удобно);

2) от формы ПЖР, имеющего форму мужского рода, оканчивающуюся на согласную (произносимую или не произносимую), причём форма женского рода оканчивается на произносимую согласную и немое "e": **heureux - heureuse - heureusement** (счастливо); **général -générale - généralement** (обычно), ***précis* - présise - précisément** (точно);

3) от формы ПМР, оканчивающейся на гласную, причём к форме женского рода прибавляется немое "e": **vrai -vraie- vraiment** (в самом деле); **aisé - aisée - aisément** (удобно); **obstiné - obstinée - obstinément** (упрямо); к этой группе относятся также **gaiement** (весело), **assidu - assidue - assidùment** (прилежно).

Вторая структура соответствует следующему правилу образования наречий:

4) от прилагательных, оканчивающихся на "ent" и "ant": **courant - courante - couramment** (бегло); **prudent - prudente - prudemment** (осторожно).

Выявление темы ЕЯ текста и контекста темы

В этой задаче текст рассматривается как совокупность предложений, предложение – как совокупность слов. Формируется словарь слов, входящих в текст. На этом этапе можно не рассматривать предлоги, союзы, местоимения, т. е. не знаменательные слова. Все слова и предложения индексируются. Каждому предложению соответствует список индексов слов словаря, которые составляют данное предложение. Для каждого слова формируется список индексов предложений, в которые оно входит. На множестве всех подмножеств индексов задаются теоретико-множественные операции пересечения, объединения и дополнения. С помощью этих операций устанавливаются отношения включения между подмножествами индексов.

Пусть j - номер предложения, word(j) – перечень выделенных слов предложения j, списки вхождений которых в предложения текста по мощности равны или больше двум, и list(word(j)) – объединенный список индексов предложений текста, в которые входят слова из word(j). Word(j) назовем **ядром предложения или темой предложения.** Предложения, соответствующие list (word(j)) - назовем **развитием темы или её контекстом**.

Для управления формированием темы используется вес группы слов, образующих тему. Тема образуется как набор слов некоторого предложения, каждое из которых имеет число вхождений в другие предложения текста большее, чем некоторое заданное минимальное число вхождений. Увеличение этого числа приводит к увеличению связности текста темы.

Другим средством увеличения связности текста темы является формирование темы как цепочки ассоциативно связанных слов, то есть каждая пара слов цепочки должна появляться, по крайней мере, не менее чем в двух предложениях текста.

Для поиска всех ассоциативных зависимостей между словами предложений текста можно использовать известные алгоритмы машинного обучения (Machine Learning) для поиска ассоциативных и импликативных зависимостей в символьных записях (данных).

Ассоциативно зависимые слова предложения часто оказываются и синтаксически зависимыми.

В общем случае тему и каждое предложение её контекста преобразуем к минимальному синтаксически связанному тексту. Таким образом, каждое предложение «сжимается» синтаксически в соответствии с его темой (пока это делается экспертом).

Пользователь выбирает тему, и тогда по выбранной теме и её контексту формируется конспект, содержащий только те предложения, которые входят в контекст темы. По одному тексту можно сгенерировать несколько различных контекстов. Для примера был взят фрагмент текста из статьи [9]. Алгоритм формирования тем и их контекстов, а также результаты работы алгоритма даны в **ПРИЛОЖЕНИИ**.

На основе выделения ядерных контекстов ЕЯ текстов можно решать с помощью методов машинного обучения и другие интересные задачи:

1. По заданным ключевым словам и тексту конструировать контекст этих слов с выделением ассоциативных связей этих ключевых слов с другими словами;
2. По уже построенному контексту находить тексты со сходным контекстом;
3. Выделяя наиболее важные слова для пользователя, составлять рефераты текстов, найденных в сети Интернет;
4. Определять степень связности контекстов;
5. Распознавать SPAM и отсеивать его в электронной почте, адаптируясь к клиенту, на основе слежения за корреспонденцией и выделения существенных признаков SPAM'а и не SPAM'а.

Список литературы

1. Марков А.А, Нагорный Н.М. // Теория алгорифмов. М.: Наука, 1984.

2. Финн В.К. Правдоподобные рассуждения в интеллектуальных системах типа ДСМ // Итоги науки и техника, серия «Информатика». М.: ВИНИТИ, 1991. №1-2. С. 8-44.

3. Kuznetsov, S. O., Obiedkov, S. A. Comparing Performance of Algorithms for Generating Concept Lattices // J. Exp. Theor. Artif. Intell., 2001. Vol. 14. No. 2-3. P. 183-216.

4. Naidenova, X. DIAGARA : An Incremental Algorithm for Inferring Implicative Rules from Examples. Parts I and II // Proceedings of KDS'2005. Sofia: FOI – Commerce, 2005. Vol. 1. P. 174 – 190.

5. Huntala, Y., Karkkainen, J., Porkka, P., and Toivonen, H. TANE: An Efficient Algorithm for Discovering Functional and Approximate Dependencies // The Computer Journal, 1999. Vol. 42. No. 2. P. 100-111.

6. Mephu Nguifo, E. and Njiwoua, P. Using Lattice Based Framework as a Tool for Feature Extraction // Feature Extraction, Construction, and Selection: A Data Mining Perspective. Lui, H. and Motoda, H., Eds. Kluwer, 1998.

7. Stumme, G., Wille, R., Wille, U. Conceptual Knowledge Discovery in Databases Using Formal Concept Analysis Methods // Proceeding The 2nd European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD'98), 1998.

8. Найденова К.А. Нормальный алгорифм Маркова как основа моделирования процессов обучения естественному языку // Обработка текста и когнитивные технологии. М.: МИСИС, 2000. Вып. 4. Часть 2. С.231-242.

9. Гладун А., Рогушина Ю. Онтологии как перспективное направление интеллектуализации поиска информации в мульти-агентных системах е-коммерции» // Труды KDS-2005. Sofia : FOI-COMMERCE, 2005. Том 1. С. 158-159.

ПРИЛОЖЕНИЕ

Алгоритм выделения контекста:

Первый этап:

по тексту строится следующая структура:

- текст разбивается на предложения (предложения индексируются - IdClause);
- предложения разбиваются на слова (слова индексируются - IdWord);
- для каждого слова формируется список индексов предложений, в которые оно входит.

На этом этапе лучше рассматривать только знаменательные слова и не рассматривать предлоги, союзы, местоимения и служебные слова (распознавание этих слов – по специальному словарю).

В результате будет сформировано множество записей следующего вида или **словарь**:

$\text{Word} \rightarrow \text{IdWord} \rightarrow \text{List (IdWord)} = \{\text{IdClause}_{k1}, \text{IdClause}_{k2}, \ldots., \text{IdClause}_{ks}\}$, $k1, k2, \ldots.ks \rightarrow \in \{1, 2, \ldots.., N\}$, где N – число предложений текста.

Образуем (с помощью эксперта или алгорифма Маркова) группы слов полученного словаря, включающие каждую словоформу и все её грамматические видоизменения, например, как в примере, данном ниже:

Group = {информационным, информационном, информационных, информационные, информационное};

Для каждой группы слов построим множество List (Group) = {$\cup$ List (IdWord), Word $\in$ Group}.

Если в тексте встречается одна словоформа, то группа будет содержать только одно слово.

Проиндексируем полученные группы, в результате будет сформировано множество записей следующего вида или список групп:

$\text{Group} \rightarrow \text{IdGroup} \rightarrow \text{List (IdGroup)} = \{\text{IdClause}_{k1}, \text{IdClause}_{k2}, \ldots., \text{IdClause}_{ks}\}$, $k1, k2, \ldots.ks \rightarrow \in \{1, 2, \ldots.., N\}$, где N – число предложений текста.

При желании эксперта в группу можно включать не только все словоформы одного слова, но и однокоренные слова, независимо от того, какой частью речи они являются. Так, в предыдущую группу может войти слово «информация». Вообще с помощью групп можно влиять на формирование тем и контекстов.

Группы можно формировать с помощью обучения по примерам.

Введем и вычислим величину Weight = $\left|\right|$ List (IdGroup) $\left|\right|$ для каждой группы, то есть число предложений, в которых встречаются слова группы с соответствующим IdGroup.

В каждом предложении j, j = {1, 2, ….., N} выберем те слова, которые входят в группы с весом Weight $\geq$ minweight (пусть minweight = 2), где minweight – минимально допустимое значение веса группы. Тогда имеем:

$\text{Word}_1 \rightarrow \text{Group}_1 \rightarrow \text{IdGroup}_1[j]$,
$\text{Word}_2 \rightarrow \text{Group}_2 \rightarrow \text{IdGroup}_2[j]$,
…….

$Word_m \rightarrow Group_m \rightarrow IdGroup_m[j]$

и

$Word_1[j] \rightarrow IdGroup_1[j]$, $IdGroup_1[j] \rightarrow List\ (IdGroup_1[j])$, $Weight = List\ (IdGroup_1[j]) \geq 2$,
$Word_2[j] \rightarrow IdGroup_2[j]$, $IdGroup_2[j] \rightarrow List\ (IdGroup_2[j])$, $Weight = List\ (IdGroup_2[j]) \geq 2$,
……
$Word_m[j] \rightarrow IdGroup_m[j]$, $IdGroup_m[j] \rightarrow List\ (IdGroup_m[j])$, $Weight = List\ (IdGroup_m[j]) \geq 2$.

Объединим выделенные слова предложения, то есть, образуем множество

KeyWord(j) = { Word1, Word2, …., Wordm }.

KeyWord(j) назовем ядром предложения или темой предложения. В частном случае тема может состоять только из одного слова, а контекст определяться только списком вхождений этого слова (группы слов) в предложения текста.

В принципе можно задавать различные значения minweight и получать разные темы в одном и том же предложении.

Образуем множество CONTEXT (KeyWord(j)) = {$\cup$ List (IdGroupi[j])}, i = {1, 2, ..., m}, где m – число слов в KeyWord(j);

Назовем CONTEXT (KeyWord(j)) развитием темы или её контекстом.

На множестве контекстов можно построить структуру по отношению сходства. Можно производить над контекстами разные операции, например, операцию объединения контекстов и операцию пересечения контекстов. Если два контекста пересекаются, то можно говорить, что соответствующие темы связаны через те предложения, номера которых попали в пересечение. Можно говорить о глубине связи тем (или контекстов).

Второй этап:

- темы упорядочиваются по длине соответствующих им контекстов, и затем можно отобрать те контексты, длина которых не меньше некоторой заданной величины;
- пользователю предъявляются те темы, которые имеют контексты, удовлетворяющие заданным требованиям по длине;
- для каждой выбранной темы предъявляются все предложения, которые входят в её контекст.

Таким образом, по одному тексту можно сгенерировать несколько различных контекстов и построить структуру связей этих контекстов.

Для примера возьмем фрагмент текста из статьи «Онтологии как перспективное направление интеллектуализации поиска информации в мультиагентных системах е-коммерции» (Анатолий Гладун и Юлия Рогушина, стр. 158-159, том 1, труды KDS-2005).

Исходный текст:

«**1.** Сегодня мы являемся свидетелями и участниками эволюции постиндустриального общества в общество, называемое информационным. **2.** В информационном обществе приоритетным направлением является создание и эффективное использование знаний и информационных ресурсов. **3.** Колоссальные перспективы развития рынка товаров и услуг в сети Интернет впечатляют даже специалистов – только за последние месяцы прошлого года торговый оборот сети возрос в несколько раз. **4.** Стремительный технический прогресс в этой области дает мощный импульс глобализации мировой экономики и делает все более привлекательным объектом инвестиций новые информационные технологии, направленные на развитие электронной коммерции (е-коммерции). **5.** Однако происходящие изменения приводят к возникновению ряда новых проблем. **6.** Поистине огромный объем предложений, широкое разнообразие товаров и услуг, высокая динамика изменений рынка – все это приводит к резкому возрастанию сложности и трудоемкости работы, как продавцов, так и покупателей в сети, отнимает их время и тем самым повышает стоимость данных услуг. **7.** Нужна кардинальная смена самой концепции обработки информации в сети Интернет, которая бы позволила более содержательно отвечать запросам клиентов, более оперативно реагировать на имеющиеся требования и гибко адаптироваться к условиям рынка. **8.** Вхождение Украины в мировое информационное пространство требует решения многоаспектной проблемы автоматизации современных бизнес-приложений. **9.** Под электронным бизнесом понимают все формы бизнес-деятельности, такие как е-коммерция, е-консалтинг, е-издательство и т.п. **10.** Е-коммерция является частным случаем электронного бизнеса. **11.** Под е-коммерцией понимают различные формы торговли товарами и услугами посредством использования электронных средств, в том числе и Интернета. **12.** При этом заказ товаров осуществляется через телекоммуникации, а расчеты между покупателем и продавцом – при помощи

электронных средств платежа. **13.** Улучшение эффективности выполнения задач е-бизнеса требует дальнейшего развития методов автоматизации бизнес-процессов. **14.** Системы е-коммерции должны обеспечивать потребителю доступ к информации о товарах, представленной в электронной форме, и её быстрый поиск в сетевой среде. **15.** Сложность транзакций очень велика из-за динамичности и огромного количества информации, доступной пользователям через Интернет. **16.** Индустриальная разработка программного обеспечения для е-коммерции требует создания и использования соответствующих моделей, стандартов, языков и форматов, ориентированных на обработку знаний. **17.** Для решения этих задач с успехом применяются агентно - ориентированные технологии, базирующиеся на использовании интеллектуальных программных агентов».

В таблице 1 приведены результаты анализа первых трех предложений.

Таблица 1. Результаты анализа первых трех предложений текста

Индекс предложения	**Группа (Group)**	**List (IdGroup)**
1	Сегодня	1
1	Мы	1
1	Являемся, является	1, 2
1	Свидетелями	1
1	Участниками	1
1	Эволюции	1
1	Постиндустриального	1
1	**Общества**, **общество**, обществе	1, 2
1	Называемое	1
1	**Информационным**, информационном, информационные, информационное, информационных	1, 2, 4, 8
---	---	---
2	Информационным, **информационном,** информационные, информационное, **информационных**	1, 2, 4, 8
2	Общества, общество, **обществе**	1, 2
2	Приоритетным	2
2	Направлением	2
2	Является	2, 10
2	**Создание**, создания	2, 16
2	Эффективное	2
2	**Использование**, использования	2, 11, 16
2	Знаний	2, 16
2	Ресурсов	2
---	---	---
3	Колоссальные	3
3	Перспективы	3
3	**Развития**, развитие,	3, 4, 13
3	Рынка	3, 6, 7
3	**Товаров,** товарами, товарах	3, 6, 11, 12, 14
3	**Услуг**, услугами	3, 6, 11
3	Сети	3, 6, 7
3	**Интернет**, Интернета	3, 7, 11, 15
3	Впечатляют	3
3	Специалистов	3
3	Только	3
3	Последние	3
3	Месяцы	3
3	Прошлого	3
3	Года	3

Индекс предложения	Группа (Group)	List (IdGroup)
3	Торговый	3
3	Оборот	3
3	Возрос	3
3	Несколько	3
3	Раз	3

Тема образуется как набор слов (с соответствующей группой слов) некоторого предложения, которые имеют число вхождений в другие предложения текста больше некоторого заданного числа.

Для первых трех предложений получились следующие наборы ключевых слов при minweight = 2:

1. являемся, общества, общество, информационным.
2. обществе, является, создание, использование, знаний, информационных.
3. развития, рынка, товаров, услуг, сети, Интернет.

Ключевые слова не образуют в общем случае связного текста. Для предъявления пользователю хорошо бы преобразовывать тему из набора ключевых слов в минимальный синтаксически связанный набор слов предложения, включающий ключевые слова или часть ключевых слов. Для этого может потребоваться добавить некоторые слова к ключевым словам (например, предлоги, союзы, к подлежащему – сказуемое) или удалить некоторые ключевые слова (пока алгоритма этого синтаксического сжатия рассматривать не будем).

С учетом синтаксической связности текста темы сформировались следующим образом:

Тема 1: Являемся свидетелями эволюции постиндустриального общества в общество, называемое информационным. **Тема 2**: Приоритетным направлением является создание и использование знаний и информационных ресурсов. **Тема 3**: Перспективы развития рынка товаров и услуг в сети Интернет.

Предложения из контекста в простейшем случае предъявляются пользователю полностью. Но также как и при формулировании темы, существует проблема минимизации предложений, то есть получения минимального синтаксически связного подпредложения, которое содержит ключевые слова.

Если выбрать **тему 3**, то получим следующий конспект, в который войдут 9 отредактированных (минимизированных) предложений 3,4,6,7.11,12,13,14,15:

«**3.** Перспективы **развития рынка товаров и услуг в сети Интернет** впечатляют специалистов – за последний месяц прошлого года торговый оборот **сети** возрос в несколько раз. **4.** Технический прогресс делает все более привлекательным объектом инвестиций новые информационные технологии, направленные на **развитие** электронной коммерции (е-коммерции). **6.** Широкое разнообразие **товаров и услуг**, высокая динамика изменений **рынка** приводит к возрастанию сложности и трудоемкости работы, как продавцов, так и покупателей в **сети**. **7.** Нужна кардинальная смена самой концепции обработки информации в **сети Интернет**. **11.** Под е-коммерцией понимают различные формы торговли **товарами и услугами** посредством использования электронных средств, в том числе **Интернета**. **12.** При этом заказ **товаров** осуществляется через телекоммуникации. **13.** Улучшение эффективности выполнения задач е-бизнеса требует дальнейшего **развития** методов автоматизации бизнес процессов. **14.** Системы е-коммерции должны обеспечивать потребителю доступ к информации о **товарах** и её быстрый поиск в сетевой среде. **15.** Сложность трансакций очень велика из-за динамичности и огромного количества информации доступной пользователю через **Интернет**».

Если величину **minweight** мы примем равной 3, то число ключевых слов, входящих в тему, несколько уменьшиться, а связность текста темы – увеличится.

Для первых трех предложений получились следующие наборы ключевых слов при **minweight = 3**:

1 информационным;
2 использование;
3 товаров, услуг, сети, Интернет.

В этом случае получим следующие темы:

Тема 1: Общество, называемое **информационным**. **Тема 2**: **Использование** знаний и информационных ресурсов. **Тема 3**: Перспективы **развития рынка товаров и услуг в сети Интернет**.

Для темы 3 контекст не изменится.

Для темы 2 получим следующий контекст:

«2. Использование знаний и информационных ресурсов. **11.** Под е-коммерцией понимают различные формы торговли товарами и услугами посредством **использования** электронных средств, в том числе и Интер-

нета. **16.** Индустриальная разработка программного обеспечения для е-коммерции требует создания и **использования** соответствующих моделей, стандартов, языков и форматов, ориентированных на обработку знаний».

Если величину **minweight** мы примем равной 4, то число ключевых слов, входящих в предложение 3, изменится:

3 товаров, Интернет.

Если выбрать **тему 3**, то получим следующий конспект, в который войдут 7 отредактированных (минимизированных) предложений 3,6,7.11,12,14,15:

«**3.** Перспективы развития рынка **товаров** и услуг в сети **Интернет** впечатляют специалистов. **6.** Широкое разнообразие **товаров** и услуг приводит к возрастанию сложности и трудоемкости работы, как продавцов, так и покупателей в сети. **7.** Нужна кардинальная смена самой концепции обработки информации в сети **Интернет**. **11.** Под е-коммерцией понимают различные формы торговли **товарами** и услугами посредством использования электронных средств, в том числе **Интернета**. **12.** При этом заказ **товаров** осуществляется через телекоммуникации. **14.** Системы е-коммерции должны обеспечивать потребителю доступ к информации о **товарах**. **15.** Сложность трансакций очень велика из-за динамичности и огромного количества информации доступной пользователю через **Интернет**».

Совершенствование процедуры выделения контекста.

Для выделения цепочек связанных (ассоциативно) слов можно производить с помощью индуктивных методов машинного обучения.

Введем понятие связности множеств некоторого семейства множеств.

Определение 1. Назовём два подмножества A, B некоторого семейства множеств M связанными, то есть A # B, если их пересечение не пусто.

Определение 2. Пусть A, C - множества некоторого семейства множеств M; A и C связаны цепью в этом семействе множеств, если существует последовательность множеств A = A1, A2, … , An = C семейства M, таких что Ai связано с Ai+1, i = 1, …, n -1.

Максимальная цепь связанных подмножеств в семействе M есть транзитивное замыкание множеств семейства по отношению #.

Каждая группа слов Group → IdGroup ассоциирована с некоторым подмножеством List (IdGroup) индексов предложений текста, в которых встречаются слова этой группы. Будем считать пару групп слов связанной, если соответствующие им подмножества индексов являются связанными подмножествами.

В каждом предложении будем искать максимальные цепи связанных слов или максимально связанные совокупности слов предложения.

Для примера возьмем предложение 3. Рассмотрим семейство подмножеств индексов предложений в столбце List (IdGroup) таблицы 1. Все подмножества семейства связаны через предложение 3, поэтому максимально связанное подмножество будет равно {3, 4, 6, 7, 11, 12, 13, 14, 15}.

Исключим индекс 3 из рассмотрения. Тогда получим следующий результат, отображенный в таблице 2:

Таблица 2. Результаты удаления индекса 3 из подмножеств индексов предложений

Индекс предложения	**Группа (Group)**	**List (IdGroup)**
3	**Развития**, развитие,	4, 13
3	Рынка	6, 7
3	**Товаров,** товарами, товарах	6, 11, 12, 14
3	**Услуг**, услугами	6, 11
3	Сети	6, 7
3	**Интернет**, Интернета	7, 11, 15

Теперь мы имеем два максимально связанных подмножества индексов предложений: {4, 13} и {3, 6, 7, 11, 12, 14, 15}.

Ведем некоторую стратификацию индексов предложений. Каждый индекс входит в конечное число подмножеств.

4 → {4, 13};
6 → {6, 7}, {6, 11}, {6, 11, 12, 14};

7 → {6, 7}, {7, 11, 15};
11 → {6, 11}, {6, 11, 12, 14}, {7, 11, 15};
12 → {6, 11, 12, 14};
13 → {4, 13};
14 → {6, 11, 12, 14};
15 → {7, 11, 15}.

Рассмотрим только такие индексы предложений, которые входят не менее, чем в два подмножества. Исключим индексы 4, 12, 13, 14, 15, тогда получим:

6 → {6, 7}, {6, 11};
7 → {6, 7}, {7, 11};
11 → {6, 11}, {7, 11}.

И максимально связанное множество индексов = {6, 7, 11}. Этому множеству соответствует набор ключевых слов: {рынка, товаров, услуг, сети, Интернет}. Теперь тема предложения 3 выглядит более компактно:

«Рынок товаров и услуг в сети Интернет»

Контекст этой темы состоит из предложений 3, 6, 7, 11:

«**3.** Перспективы **развития рынка товаров и услуг в сети Интернет** впечатляют специалистов – за последний месяц прошлого года торговый оборот **сети** возрос в несколько раз. **6.** Широкое разнообразие **товаров и услуг**, высокая динамика изменений **рынка** приводит к возрастанию сложности и трудоемкости работы, как продавцов, так и покупателей в **сети**. **7.** Нужна кардинальная смена самой концепции обработки информации в **сети Интернет**. **11.** Под е-коммерцией понимают различные формы торговли **товарами и услугами** посредством использования электронных средств, в том числе **Интернета**». Ниже мы приводим алгоритм построения максимально связанных подмножеств некоторого семейства множеств.

Алгоритм построения максимально связанных подмножеств семейства множеств:

$M = \{1, 2, \ldots N\}$ - множество индексов;

$S = \{s_1, s_2, \ldots s_d\}$ - семейство подмножеств множества M;

1. Упорядочить лексикографически множества семейства;

2. Объединить все подмножества семейства, начинающиеся с одинакового индекса. Пусть в результате образовано семейство множеств $Snew = \{ s_{1new}, s_{2new}, \ldots s_{dnew} \}$ и $SM = \cup s_{inew}$;

3. для каждого подмножества s_{inew} построить его максимальное расширение:
   3.1 формируем множество индексов CAND = $\{SM / s_{inew}\}$;
   3.2 формируем $s_{inew} = \{ s_{inew} \cup j \}$ для всех $j, j \in$ CAND, если $(\exists l)$, $l \in s_{inew}$ такое что $\{ l, j \} \subseteq s_{knew}$, $k \neq i$, $k > i$, and i - первый индекс в s_{inew} k - первый индекс в s_{knew} ;
   3.3 удалить s_{knew}, $k > i$, если $s_{knew} \subseteq s_{inew}$;

4. stop если нет более подмножеств для расширения.