О ВОЗМОЖНОСТЯХ АВТОМАТИЗАЦИИ ВЫЯВЛЕНИЯ СВЯЗЕЙ МЕЖДУ ТЕРМИНАМИ ПРЕДМЕТНОЙ ОБЛАСТИ

(НА ПРИМЕРЕ КАТАЛИЗА)

Саломатина Н.В., Гусев В.Д.
Институт математики СО РАН
Ильина Л.Ю.*, *Кузьмин А.О., Пармон В.Н.
Институт катализа
г. Новосибирск

Решаемая задача

- Формирование и пополнение специализированного тезауруса на основе текстов предметной области
- Тезаурус – термины ПО + связи между ними
- Предметная область – катализ
- Назначение – расширение и уточнение поисковых запросов, повышение эффективности поиска
- Первоначальное наполнение – из предметных указателей учебников
- Пополнение – из периодических изданий

Выявление терминов в тексте

Элементы технологии:

- формирование текстовой подборки
- выбор системы представления текстов
- обработка текстов и формирование словаря терминов и индикаторов связи:
  - с использованием процедур фильтрации и упорядочения
  - с привлечением эксперта на заключительном этапе

Представление текста

L-грамма – цепочка из L подряд следующих слов текста T

$$\Phi_L(T) = \{\varphi_{L1}(T), \varphi_{L2}(T), ..., \varphi_{LM_L}(T)\}$$ –

L-граммная характеристика текста $(L = 1, \ldots, L_{max})$

L_{max} – длина максимального повтора в T

M_L – число различных L-грамм в T

$\varphi_{Li}(T)(1 \le i \le M_L)$: x_i – i-я L-грамма

$F(x_i)$ – частота встречаемости x_i в T

$\{r_j(x_i)\}$ – позиции вхождения x_i в T

$1 \le j \le F(x_i)$

$\Phi(T) = \{\Phi_1(T), \Phi_2(T), ..., \Phi_{L_{max}}(T)\}$ – полный L-граммный спектр T

Представление группы текстов

L-граммная характеристика группы текстов $T = \{T_1, T_2, \ldots, T_m\}$

$$\Phi_L(T) = \{\varphi_{L1}(T), \varphi_{L2}(T), \ldots, \varphi_{LM_L}(T)\}$$

$\varphi_{Li}(T)$ – четверка: x_i – i-я L-грамма;

$F_T(x_i)$ – текстовая частота (число текстов из T, в которых представлена x_i);

$F_a(x_i)$ – абсолютная частота встречаемости x_i в T;

$f(x_i) = (f_1(x_i) f_2(x_i) \ldots f_m(x_i))$ – вектор частот вхождения L-граммы x_i в каждый из текстов подборки T.

Совместный L-граммный спектр группы текстов:

$$\Phi(T) = \{\Phi_1(T), \Phi_2(T), \ldots, \Phi_{L_{\max}}(T)\}$$

Схема обработки обучающей подборки

- нормализация слов
- вычисление *L*-граммных характеристик
  (на основе trie-структур)
- выявление устойчивых (встречающихся в большом числе разнообразных контекстов) цепочек (*L*-грамм)
- введение ограничений на параметры, характеризующие *L*-грамму
- уточнение терминологического словаря экспертом

Выявление связей

- Исследование вариативности *L*-грамм
  - *L*-граммные шаблоны: построение образцов с переменной/переменными
- Выявление позиционной неравномерности распределения *L*-грамм в тексте
  - построение профиля кластеризуемости *L*-грамм в тексте
- Использование индикаторов связи
  - Формирование индикаторного словаря экспертом на основе устойчивых *L*-грамм небольшой длины, не являющихся терминами
  - поиск индикаторов в тексте, анализ контекста

Терминологические шаблоны

- образец с одной переменной

$$p = a_1 a_2 \ldots a_{k-1}\ x_k\ a_{k+1} \ldots a_n$$

фиксирует подмножество словосочетаний длины n, отличающихся друг от друга заменой по k-й позиции

Пример

p = **РЕАКТОР\С\X\СЛОЕМ**

X $\in$ **{ НЕПОДВИЖНЫМ, КИПЯЩИМ, ДВИЖУЩИЙМСЯ, ПСЕВДОСЖИЖЕННЫМ, ФИЛЬТРУЮЩИМ }**

Объединение образцов **X\КАТАЛИЗАТОР\, КАТАЛИЗАТОР\Y**

X ∈ { активность, состав, приготовление, восстановление, свойство, селективность, …}

Y ∈ { окисление, крекинг, гидрирование, полимеризация, газоочистка,…}

Образец p = **X\КАТАЛИЗАТОР\Y** позволяет учесть всевозможные комбинации слов:

АКТИВНОСТЬ КАТАЛИЗАТОРА ОКИСЛЕНИЯ
АКТИВНОСТЬ КАТАЛИЗАТОРА КРЕКИНГА
СОСТАВ КАТАЛИЗАТОРА ОКИСЛЕНИЯ
СОСТАВ КАТАЛИЗАТОРА КРЕКИНГА
ПРИГОТОВЛЕНИЕ КАТАЛИЗАТОРА ГАЗООЧИСТКИ
ВОССТАНОВЛЕНИЕ КАТАЛИЗАТОРА ОКИСЛЕНИЯ
СВОЙСТВО КАТАЛИЗАТОРА ОКИСЛЕНИЯ
СЕЛЕКТИВНОСТЬ КАТАЛИЗАТОРА ОКИСЛЕНИЯ

Профиль позиционной кластеризуемости терминов в тексте

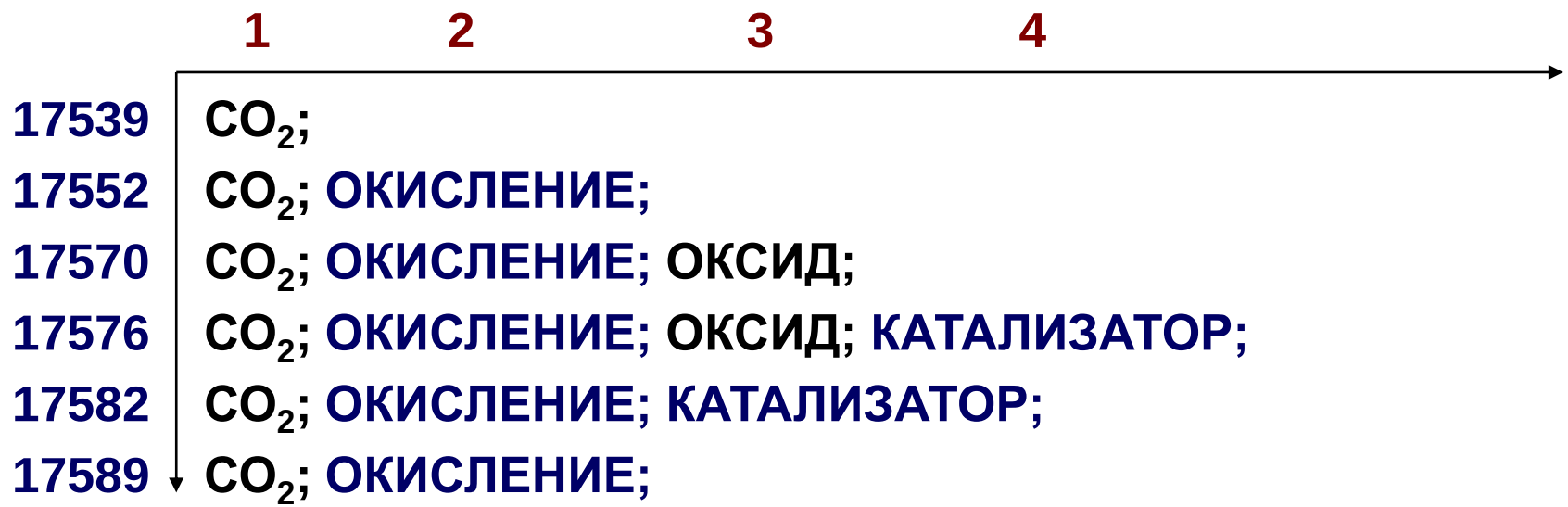
- Статистически значимые кластеры выделяются с помощью сканирующих статистик
- Взаимное расположение кластеров:
  - позиционно разнесены друг от друга
  - пересекаются друг с другом
  - вкладываются один в другой
- Профиль кластеризуемости аккумулирует на одном графике информацию обо всех участках кластеризации разных *L*-грамм

Профиль кластеризуемости терминов в тексте

– ступенчатая функция

аргумент – порядковый номер предложения в тексте

значение – число различных кластеров,
включающих в себя данное предложение



Индикаторы связи

«Он (Ипатьев)… создал ряд важнейших каталитических процессов нефтепереработки, таких как алкилирование, гидрокрекинг, изомеризация»

- Индикаторы связи отбираются экспертом из устойчивых цепочек параллельно с формированием словаря терминов
- Индикаторный подход прост в реализации, результаты хорошо интерпретируемы
- Ограничения подхода:
  - необходимость формирования индикаторных словарей в каждом отдельном случае
  - отсутствие гарантий обязательного наличия индикатора

Индикаторы связи

- поиск подстроки (индикатора) в строке (тексте):
  - Бойер-Мур – $O(N)$
  - Trie-структуры \ L-граммные деревья – $O(N \bullet L)$
  - Кнут-Моррис-Пратт – $O(N)$
  - Ахо-Корасик (групповой запрос) – $O(N + \Sigma|p_i|)$

N – длина текста

$|p_i|$ – длина образца

Структура тезауруса

- Русскоязычная часть тезауруса – 1035 терминов
- NT – narrower term: *метанол, окисление* – *метанол*
- BT – broader term: *окисление каталитическое* – *метанол*, *окисление*
- USE – use instead: *метанол* – *древесный спирт*
- UF – use for: *древесный спирт* – *метанол*
- RT – related term: *метанол*, *окисление* – *кислород*
- LE – linguistic equivalent: *метанол* – *methanol*
- x-FE – full equivalent:

  метанол, *окисление в формальдегид* – *формальдегид*, *получение окислением метанола*

Количественные характеристики тезауруса

- В среднем ~ 3 связи на термин
- Максимальное число связей – 26
  катализаторы окисления – 26, переходные металлы – 20
- Минимальное число связей – 0
  стехиометрическое число, оже-спектроскопия, Рейнольдса критерий,…
- Распределение количества связей у термина

1	2	3	4	5	6	7	8	9	10	≥ 10
7%	19%	21%	14%	10%	6%	5%	4%	4%	1%	9%

- Распределение связей по типам:

NT	BT	LE	RT	x-FE	USE,UF
33%	32%	16%	10%	8%	1%

Апробация методов

- Тексты:

1. О.В. Крылов «Гетерогенный катализ»
2. В.Б. Фенелонов «Введение в основы адсорбции и текстурологии»
3. И.П. Мухленов «Технология катализаторов»
4. «Лекции по катализу»
5. «Химическая энциклопедия»

Количественные характеристики текста

- Объем текста ~ 400 тыс. словоупотреблений
- Объем словаря текста ~ 24 тыс. слов
- Устойчивые *L*-граммы:

L = 2	3	4	5	
14	3,8	0,6	0,09	тыс.

- Расслоение лексики − по с.к.о.

  начало списка содержит 80 − 100 % терминов

  конец списка − 10 − 15 %

Образцы

F = 250 активные центры

0	LE active centers	
2	NT активные центры Бренстеда	+
0	NT активные центры, кислотные	+
0	NT активные центры Льюиса	+
0	NT активные центры, основные	+
0	NT активные центры, функция распределения	
26	NT активные центры, число	+
0	RT поверхность, неоднородной поверхности теория	

X\ЦЕНТР, X ∈{ АКТИВНЫЙ_250;КИСЛОТНЫЙ_96;ОСНОВНЫЙ_24;ЧИСЛО_18…

ЦЕНТР\X, X ∈{ ЛЬЮИСА_34;БРЕНСТЕД_29;ПОВЕРХНОСТЬ_23…

X\АКТИВНЫЙ\ЦЕНТР, X ∈{ ЧИСЛО_26, КОНЦЕНТРАЦИЯ_9,…}

ЧИСЛО\X\ЦЕНТР, X ∈{ АКТИВНЫЙ_26;КИСЛОТНЫЙ_6;ОСНОВНЫЙ_3…}

X\ЦЕНТР\БРЕНСТЕД , X ∈{ КИСЛОТНЫЙ_15;ЧИСЛО_2;АКТИВНЫЙ_2…}

АКТИВНЫЙ\ЦЕНТР\X, X ∈{ ПОВЕРХНОСТЬ_9;БРЕНСТЕДА_2…}

КИСЛОТНЫЙ\ЦЕНТР\X, X ∈{ ЛЬЮИСА_16;БРЕНСТЕД_15…}

Образцы

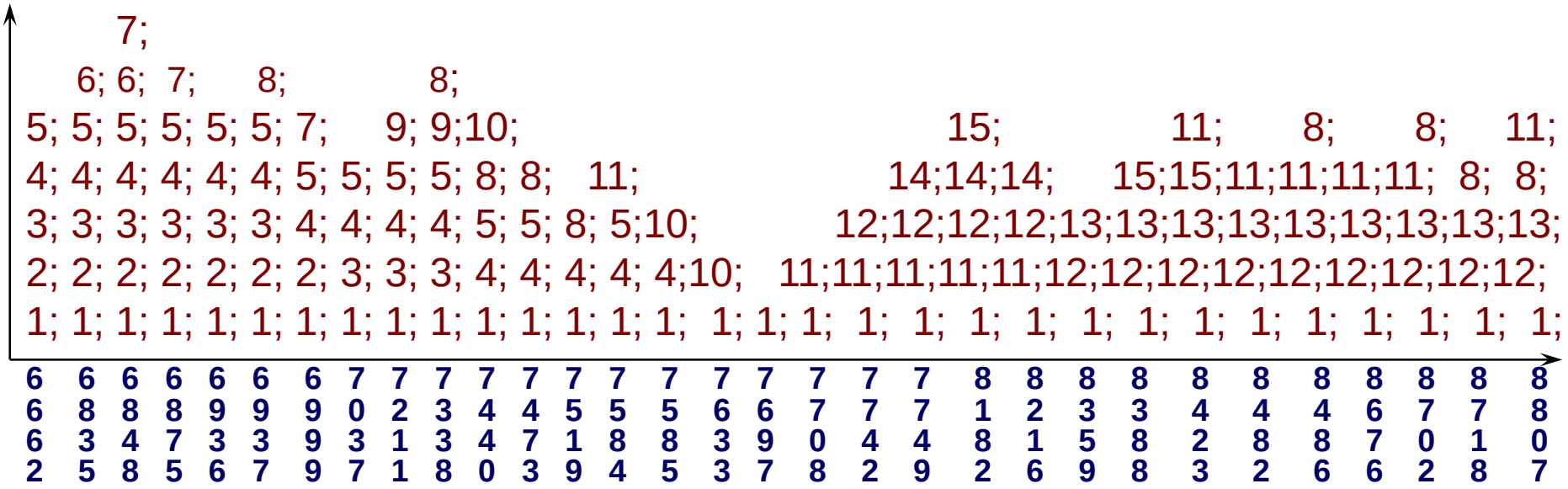
- x-FE: ОКИСЛЕНИЕ\X\B\ЭТИЛЕНОКСИД, X ∈{ЭТИЛЕН _11; C_2H_4_1 }
- RT: КАТАЛИЗАТОР\X, X ∈{полимеризация_15;Циглер-Натта_13
- NT\BT, x-FE, RT: X\УГЛЕВОДОРОД, X ∈{

1) воскообразный_4;газообразный _4;жидкий(3);твердый_2
2) высший_5; разветвленный_5; насыщенный_4; ненасыщенный_2; предельный_2; непредельный_2…
3) ароматический_57; нафтеновый_10; парафиновый_5; ацетиленовый_3; изопарафиновый_2…
4) производство_4; выход_4; переработка_2…
5) образование_9; превращение_9; реакция_4; взаимодействие_2…
6) окисление_22; синтез_17; крекинг_14; адсорбция_6; дегидрирование_5; изомеризация_3; алкилирование_2…}

Профиль кластеризуемости

Пример

1. КИСЛОТНЫЙ ЦЕНТР;
2. УДЕЛЬНАЯ ПОВЕРХНОСТЬ;
3. ПОРИСТАЯ СТРУКТУРА;
4. ОСНОВНЫЙ ЦЕНТР;
5. КАТАЛИЗАТОР КРЕКИНГА;
6. АКТИВНЫЙ ЦЕНТР;
7. АДСОРБИРОВАННАЯ МОЛЕКУЛА;
8. КАТАЛИТИЧЕСКИЙ АКТИВНОСТЬ;
9. ЦЕОЛИТ ТИПА;
10. АРОМАТИЧЕСКИЙ УГЛЕВОДОРОД;
11. ПЕРЕХОДНЫЙ МЕТАЛЛ;
12. КРИСТАЛЛИЧЕСКОЕ ПОЛЕ;
13. ИОН МЕТАЛЛА;
14. АТОМ МЕТАЛЛА;
15. ТВЕРДОЕ ТЕЛО;



Профиль кластеризуемости

активные центры

LE active centers
NT активные центры Бренстеда +
NT активные центры, кислотные +
NT активные центры Льюиса +
NT активные центры, основные +
NT активные центры, функция распределения
NT активные центры, число +
RT поверхность, неоднородной поверхности теория

L = 2;	№ фраз
АКТИВНЫЙ ЦЕНТР	6835 ÷ 6874
КИСЛОТНЫЙ ЦЕНТР	6662 ÷ 8809
ОСНОВНЫЙ ЦЕНТР	6658 ÷ 7632

L = 1; АДСОРБЦИЯ 6742 ÷ 6752, 6789 ÷ 6804

L = 3;	№ фраз
КИСЛОТНЫЙ ЦЕНТР БРЕНСТЕДА	6767 ÷ 7573
КИСЛОТНЫЙ ЦЕНТР ЛЬЮИСА	6751 ÷ 7609
ЧИСЛО АКТИВНЫХ ЦЕНТРОВ	6750 ÷ 6874
СИЛА КИСЛОТНОГО ЦЕНТРА	6662 ÷ 6776

Индикаторы связи

- Объем словаря ~ 220 индикаторов
- BT/NT

  К ПРОЦЕССАМ ГОМОГЕННОГО КАТАЛИЗА ОТНОСЯТ МНОГОЧИСЛЕННЫЕ РЕАКЦИИ ГИДРАТАЦИИ, ГИДРОЛИЗА, СУЛЬФИРОВАНИЯ, ГАЛОГЕНИРОВАНИЯ, ЭТЕРИФИКАЦИИ, КОНДЕНСАЦИИ И ДРУГИЕ

- RT

С ПОМОЩЬЮ КРЕКИНГА ИЗ НЕФТИ ПОЛУЧАЕТСЯ ЖИДКОЕ МОТОРНОЕ ТОПЛИВО: БЕНЗИН, ДИЗЕЛЬНОЕ И РЕАКТИВНОЕ ТОПЛИВО

- x-FE

ПРОМОТОРАМИ, ИЛИ АКТИВАТОРАМИ, НАЗЫВАЮТ ВЕЩЕСТВА, ДОБАВЛЕНИЕ КОТОРЫХ К КАТАЛИЗАТОРУ УВЕЛИЧИВАЕТ ЕГО АКТИВНОСТЬ, СЕЛЕКТИВНОСТЬ, УСТОЙЧИВОСТЬ.

- USE/UF

КАТАЛИТИЧЕСКОЕ ГИДРИРОВАНИЕ ИЛИ ГИДРОГЕНИЗАЦИЯ ВКЛЮЧАЕТ БОЛЬШУЮ ГРУППУ РЕАКЦИЙ ПРИСОЕДИНЕНИЯ ВОДОРОДА ПО НЕНАСЫЩЕННЫМ СВЯЗЯМ…

Точность поиска связанных терминов

- один из : найдено 180 фраз, из них верно – 87, p = 48%
- один из … являться : 43/34, p = 79%
- и другой : 229/169, p = 74%
- и др. : 284/236, p = 83%
- синтез … из …
- окисление … в …

Способ повышения точности – построение комбинированных индикаторов:

а) индикатор + индикатор
б) индикатор + термин
в) построение образцов

Заключение

- Предложены три возможные подхода, которые позволяют выявлять из текстов термины, связанные зафиксированными в тезаурусе отношениями.
- Рассмотрены возможности частичной автоматизации процесса выявления связей, ориентированные на минимизацию труда эксперта
- С помощью предложенных методов обнаруживаются:
  - связи, отсутствующие в текущей версии тезауруса,
  - новые термины, связанные с имеющимися в тезаурусе
- Дублирование найденных разными методами связей может служить подтверждением правильности выявленной связи

Поиск устойчивых цепочек

a, b – словоформы, x_L – L-грамма, $F(x_L)$ – ее частота в T

a^*x_L, x_Lb^* – лево- и правосторонние расширения x_L с максимальными $F(ax_L)$ и $F(x_Lb)$.

Критерий устойчивости: x_L с $F(x_L) > 2$ устойчива, если $F(a^*x_L) / F(x_L) \le \Pi$ и $F(x_Lb^*) / F(x_L) \le \Pi$

устойчивые сочетания	**неустойчивое сочетание**
предложение с	предложение со
предложение со Скремблингом	

Позиционный анализ

- Тематически важные слова распределены по тексту неравномерно
- Типы неравномерности:

кластеры, гэпы, сверхравномерное распределение

- Способы выявления неравномерности:

сканирующие статистики, с.к.о., …

Выявление кластеров устойчивых цепочек в тексте

Метод выявления – обнаружение аномалий в позиционном распределении ЯЕ

Аппарат – сканирующие статистики

$d(n)$ – минимальный размер интервала d с фиксированным числом (n) вхождения ЯЕ (вычисляется для каждой ЯЕ из *T)*

Если $d(n)$ – аномально мал, то ЯЕ кластеризуются

Значимость кластера оценивается с помощью имитационного моделирования

Выявление кластеров устойчивых цепочек в тексте

Кластеризация имеет место, если выполняется условие:

$$(S_{набл} \le S_{min}) \& (S_{набл} \le \bar{S} - 3s)$$

$S_{набл}$ $\bar{S}$ – наблюдаемое значение $d(n)$ в тексте

S_{min} и – минимальное и среднее значения $d(n)$ в имитационном эксперименте

s – среднеквадратичное отклонение

- Микротема характеризуется кластерами, содержащими по 6÷12 вхождений устойчивых цепочек

- Среднее внутрикластерное расстояние между устойчивыми цепочеками меньше среднего внутритекстового в 5 и более раз