Эффективный поиск нечетких заимствований

Аникеев Дмитрий Александрович

Научный руководитель: к.ф.-м.н. Ю.В. Чехович
Московский физико-технический институт
Факультут управления и прикладной математики
Кафедра интеллектуальных систем

13 июня 2018 г.

Мотивация

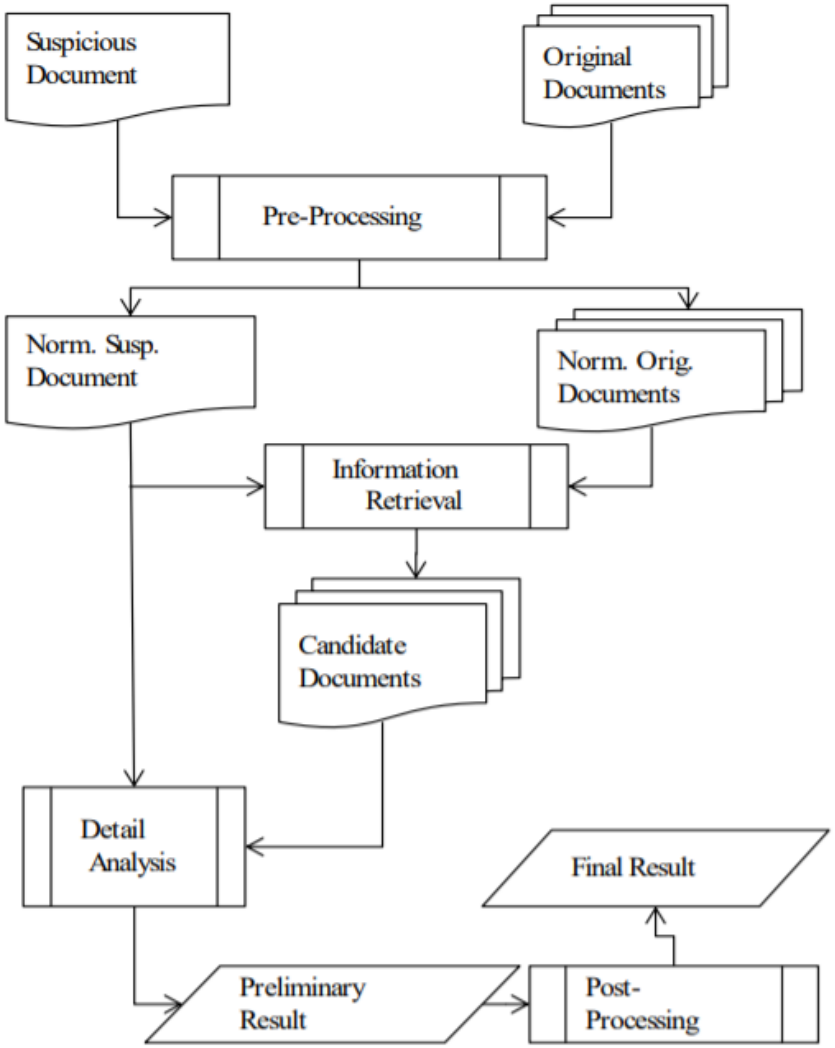
Проблема

Существует зазор в быстродействии и качестве между четким поиском и семантическими(контекстными) моделями.

Цель работы

- Промежуточный алгоритм между алгоритмами решения задачи парафраза и четким поиском дубликатов.
- Алгоритм, требующий малую вычислительную мощность.
- Алгоритм, демонстрирующий адекватное значение качества поиска заимствований.

Используемая модель



Постановка задачи четкого поиска

Терминология

Назовем $\Omega = [\omega_1, \cdots, \omega_n]$ - исходным документом, а $S = [s_1, \cdots, s_m]$ - источником. $s_i, \omega_i \in \Sigma$.
Назовем $Dup : \mathbb{N}^3 \to \{0, 1\}$ такую, что

$$Dup(l_1, l_2, length) = 1 \Leftrightarrow \forall j \in [0, length - 1] \;\; \omega_{l_1+j} = s_{l_2+j}$$

$$maxDup : \mathbb{N}^3 \to \{0, 1\}$$

$$maxDup(l_1, l_2, length) = 1 \Leftrightarrow \begin{cases} Dup(l_1, l_2, length + 1) = 0 \\ Dup(l_1, l_2, length) = 1 \\ Dup(l_1 - 1, l_2 - 1, length + 1) = 0 \end{cases}$$

Задача

Найти $(max)Dup^{-1}(1) \cap \{length \geq length_{min}\}$

Формальная постановка задачи

Дана функция расстояния $\rho(q, r) : \Sigma^* \mapsto R$.

$fuzzyDup(l_1, l_2, length_1, length_2) = 1 \Leftrightarrow$

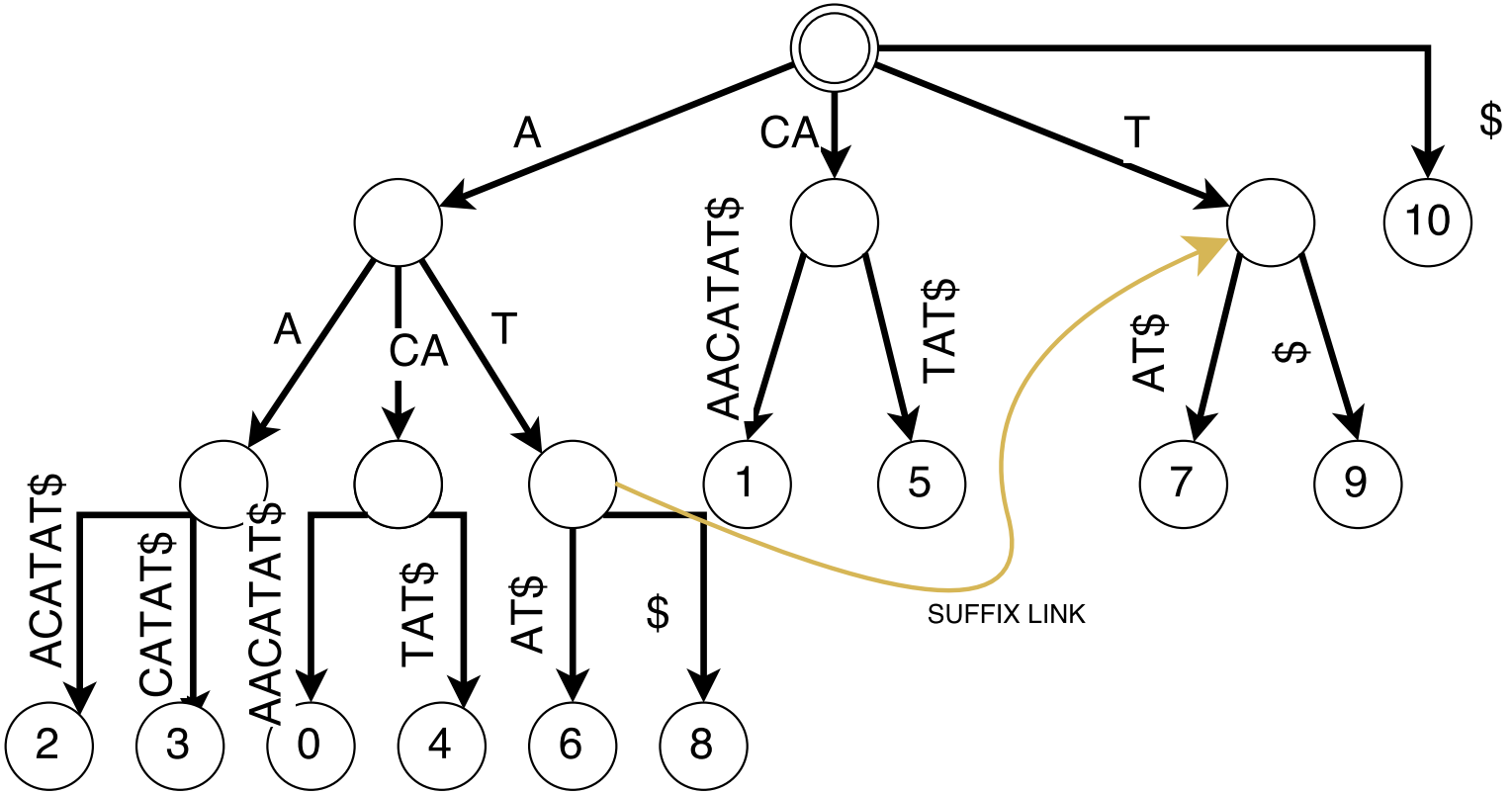
$\rho(\omega_{[l_1, l_1+length_1]}, s_{[l_2, l_2+length_2]}) \underbrace{\leq \quad \varepsilon}_{\text{Внешний критерий}}.$

$$maxfuzzyDup(l_1, l_2, length_1, length_2) = 1 \Leftrightarrow \begin{cases} Dup(l_1, l_2, length) = 1 \\ \text{Нельзя расширить строку до другой нечеткой дубликата} \end{cases}$$

Задача

Найти $(max)FuzzyDup)^{-1}(1) \cap \{length_1 > minlength\}$

Базовый алгоритм



Dong Kyue Kim, Minhwan Kim, Heejin Park. **Linearized Suffix Tree.** *Algorithmica*, 2007.

Mohamed Ibrahim Abouelhoda, Stefan Kurtz,Enno Ohlebusch. **Replacing suffix trees with enhanced suffix arrays** *Journal of Discrete Algorithms*, 2004.

Предложенный алгоритм

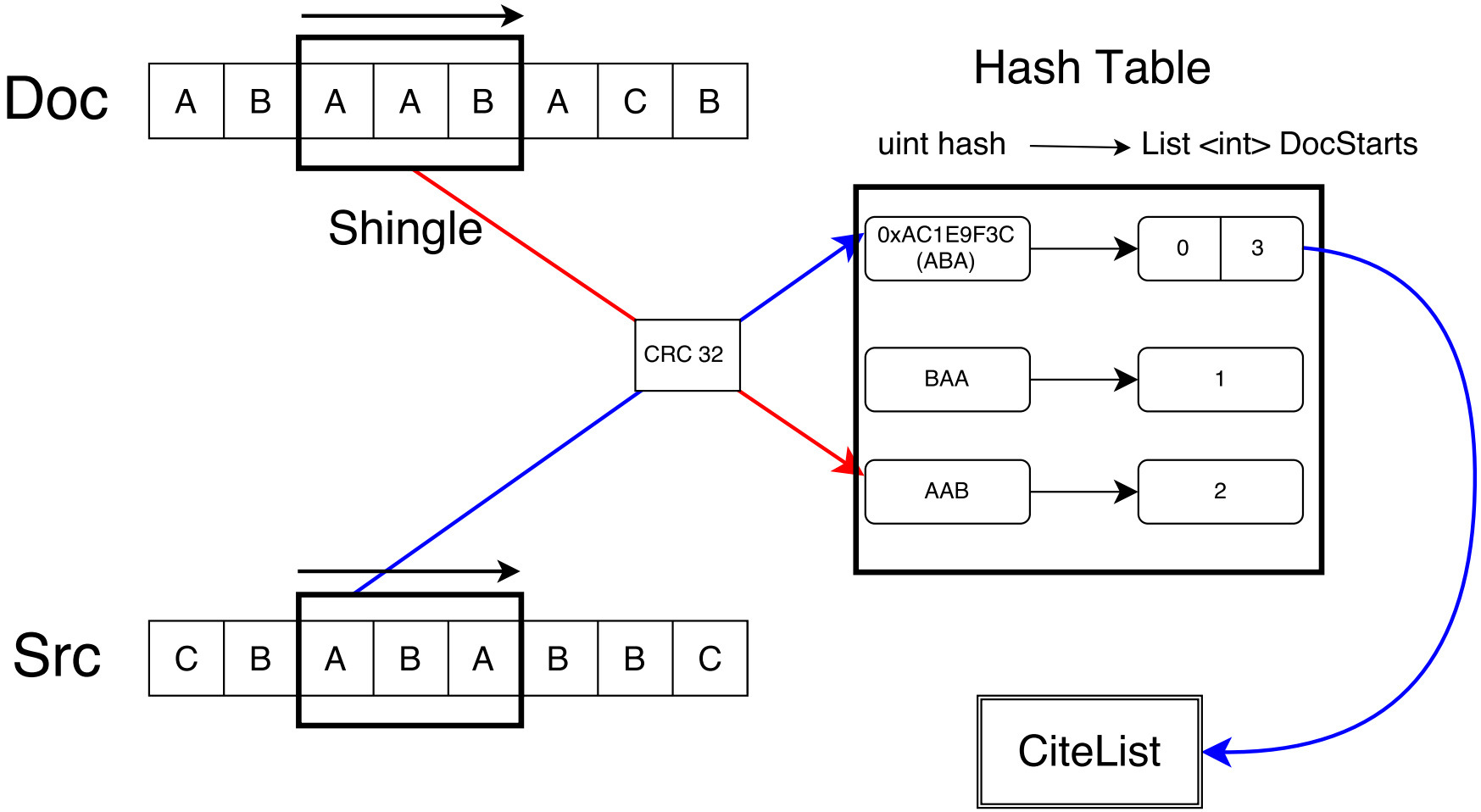
Алгоритм

- **Фаза 1** Определение коротких четких дубликатов.
- **Фаза 2** Процедура распространения четких дубликатов.
- **Постобработка** Склейка дубликатов.

Вычисление расстояний Левенштейна

$$\rho(m,n) = \min \begin{cases} \rho(m-1,n)+1 \\ \rho(m,n-1)+1 \\ \rho(m-1,n-1)+(int)(doc[m]! = doc[n]) \end{cases}$$

Фаза 1


Doc
A B A A B A C B
Shingle
Hash Table
uint hash
List <int> DocStarts
0xAC1E9F3C
(ABA)
0 3
BAA
1
AAB
2
CRC 32
Src
C B A B A B B C
CiteList

Расстояние Левенштейна

Расстояние Левенштейна — минимальное необходимое количество замен/удалений/вставок для приведения одной строки к другой.

	A	B	C	D	E	F	G	H
A	0	1	2	3	4	5	6	7
C	1	1	1	2	3	4	5	6
C	2	2	1	2	3	4	5	6
D	3	3	2	2	3	4	5	6
F	4	4	3	3	3	3	4	5
E	5	5	4	4	3	4	4	5
F	6	6	5	5	3	3	4	5
G	7	7	6	6	4	4	3	4

$$D(i,j) = \begin{cases} i, & j = 0 \\ j, & i = 0 \\ min\{ & \\ \quad D(i,j-1)+1 & \\ \quad D(i-1,j)+1 & \texttt{иначе} \\ \quad D(i,j)+(int)(doc[i] != src[j]) & \\ \quad \}, & \end{cases}$$

Рис.: Демонстрация вычисления расстояния Левенштейна на отсечении Укконена.

Задача поиска оптимальных параметров модели

PlagEvalRus

Есть выборка $\mathfrak{D} = \left\{\{doc_i, src_i\} \to \{List\ <Dups>\ TrueDups_i\}\right\}$
Задано отображение $Quality(DupList, TrueList)$:
$List<Dups>^2 \mapsto [0, 1]$, вычисляющее близость разметок документов с помощью микро-усредненной точности, полноты, и др.

Potthast M., Stein B., Barron-Cedeno A., Paolo R. **An Evaluation Framework for Plagiarism Detection** *23rd International Conference on Computational Linguistics, COLING 2010.*

Задача оптимизации

Дан параметризованный алгоритм $A(\mathbf{w}) : (\Sigma^*)^2 \mapsto List<Dups>$
Необходимо решить задачу
$$\hat{\mathbf{w}} = \arg\max_w \sum_{i \in [1, |\mathfrak{D}|]} \frac{1}{|\mathfrak{D}|} Quality(A(w, doc_i, src_i), \mathfrak{D}(doc_i, src_i))$$

Параметры алгоритма

Числа

- MinCiteLength
- LIMIT
- EXPAND
- GLUE

Процедуры

- Склейка
- Лемматизация
- Стоп-слова
- Исправление опечаток

Группа	Заимствования
gen copy(4250 пар)	Дословные
gen par(4250 пар)	Слабо и средне модифицированные
man par (713 пар)	Все виды заимствований
man par2 (198 пар)	Средне и сильно модифицированные

Метрики качества: Precision, Recall, Granularity, PlagDet.

$$PlagDet = \frac{F_1}{\ln_2(1 + Granularity)}$$

Функционал качества

$$Q_{prec} = \frac{|\bigcup\limits_{i,j,i',j'}(Dups_{i,j} \sqcap TD_{i',j'})|}{|\bigcup\limits_{i',j'} TD_{i',j'}|}$$

$$Q_{recall} = \frac{|\bigcup\limits_{i,j,i',j'}(Dups_{i,j} \sqcap TD_{i',j'})|}{|\bigcup\limits_{i,j} Dups_{i,j}|}$$

Гранулярность разметки документов

$$gran(Dups, TD) = \frac{1}{|Dups_{TD}|} \sum_{dup \in Dups_{TD}} |TD_{dup}|,$$

где $Dups_{TD} = \{Dups_j \in Dups : \quad \exists j' : TD'_j \cap Dups_j \neq \emptyset\}$, а $TD_{dup} = \{td \in TD : \quad td \cap dup \neq \emptyset\}$.

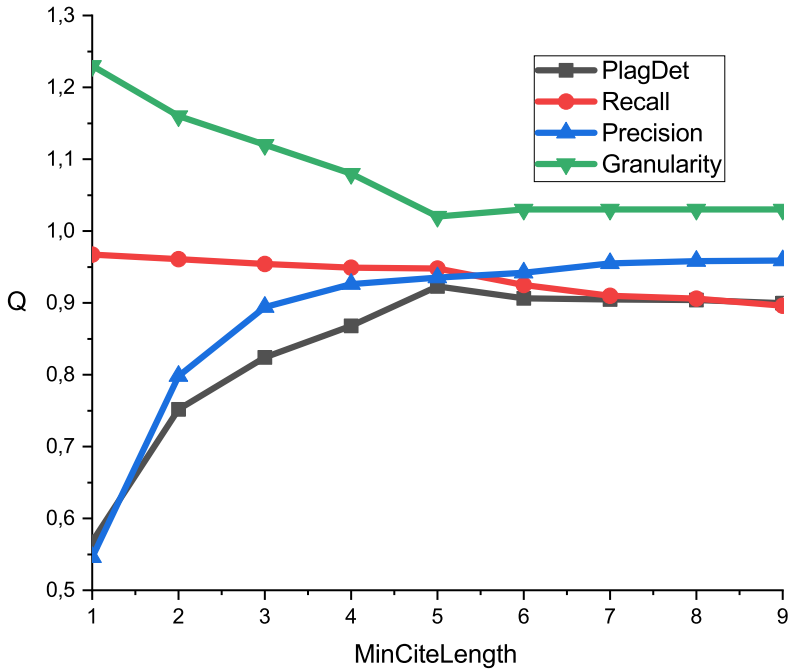
Таб. 1. Оптимальные параметры модели

Параметр	Manually_Paraphrased	Manually_Paraphrased 2
MinCiteLength	5	4
LIMIT	5	6
EXPAND	2	1
GLUE	16	23

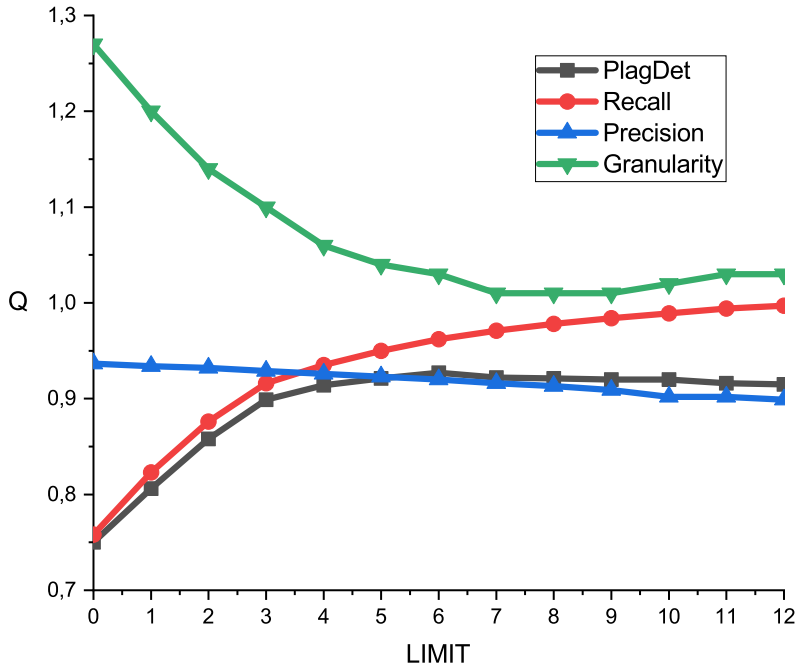
Таб. 2. Качество модели нечеткого поиска на различных коллекциях

Метрика	Gen_Copy	Gen_Para	Man_Para	Man_Para 2
Precision	0.932	0.983	0.936	0.854
Recall	0.997	0.967	0.905	0.586
Granularity	1.001	1.001	1.008	1.013
PlagDet	0.961	0.780	0.922	0.551

Влияние параметров модели на распознавание



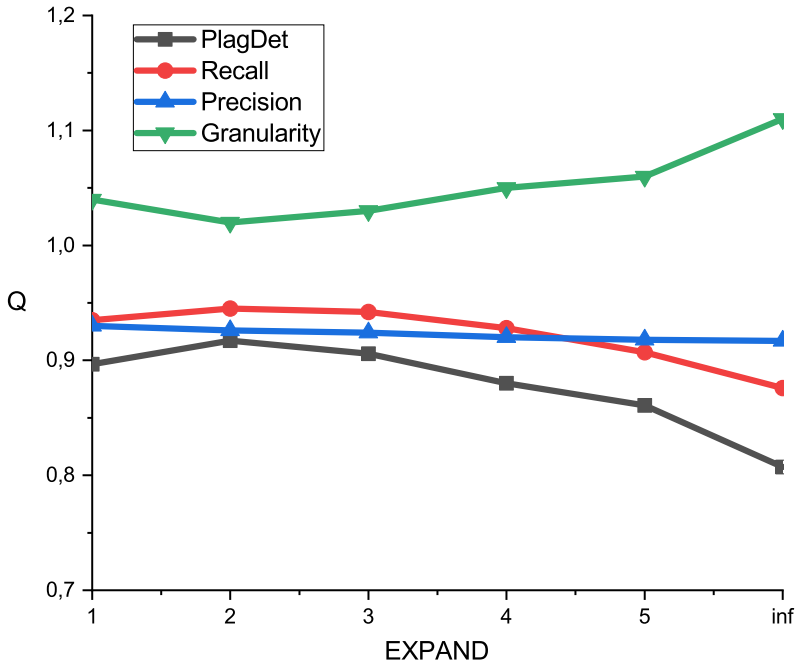
(a) MinCiteLength



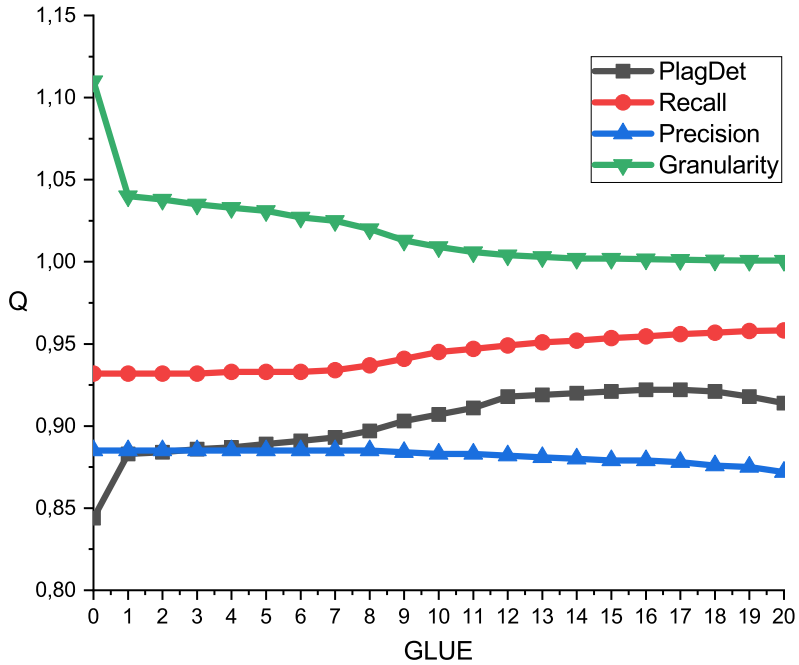
(b) LIMIT

Рис.: Зависимость качества обнаружения заимствований от параметров модели

Влияние параметров модели на распознавание



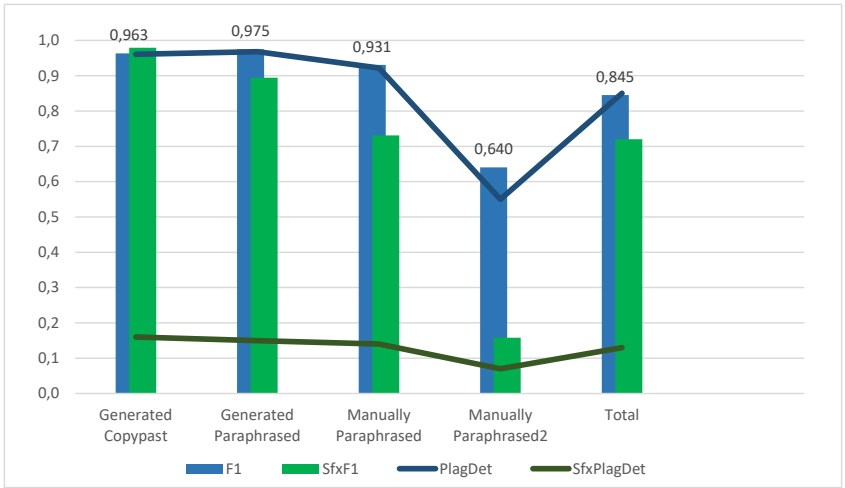
(a) EXPAND



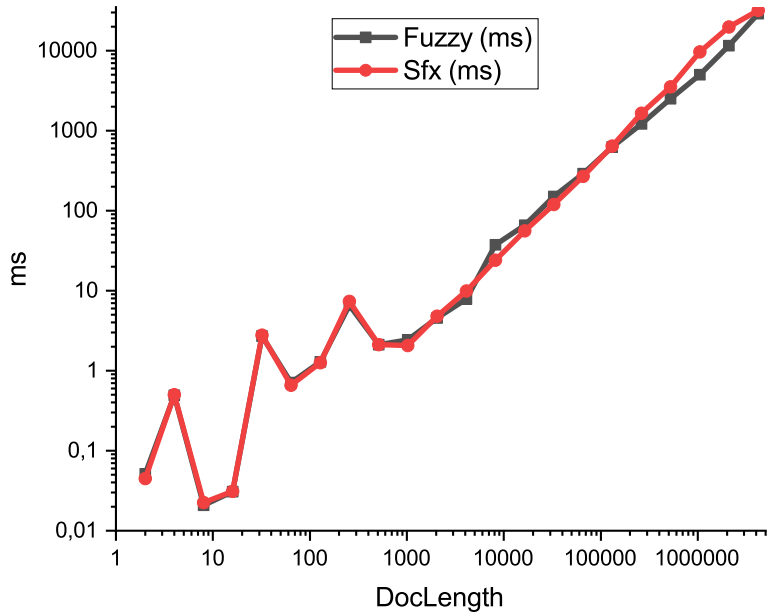
(b) GLUE

Рис.: Зависимость качества обнаружения заимствований от параметров модели

Качество и быстродействие алгоритмов



(a) Качество



(b) Производительность

Рис.: Сравнение качества и быстродействия моделей

Виды заимствований manually Paraphrased

Par Type	SfxArray		FuzzySearch	ParplagDet
Type	Recall	Granularity	Recall	Recall
CPY	98,6	13,1	99,5	95,6
ADD	61,1	9,57	96,4	93,9
DEL	63,6	10,9	95,9	93,7
CCT	88,2	17,0	98,8	96,6
SSP	87,2	19,4	98,5	96,8
LPR	35,6	7,67	94,1	94,2
HPR	11,5	6,24	62,3	78,6

Зависимость полноты от типа заимствования.

Hamza Osman A., Salim N., Salem M., Alteeb R., Abuobieda A. **An Improved Plagiarism Detection Scheme Based On Semantic Role Labeling** // Applied Soft Computing, 2012.

Выводы

- Нечеткий поиск показывает ожидаемое качество распознавания заимствований с небольшими и среднеми модификациями, значительно выше суффиксного массива, однако уступает решениям задачи поиска парафраза на коллекции с сильными модификациями.
- Алгоритм нечеткого поиска сравним по быстродействию с суффиксным массивом. Ассимптотическая сложность $O(|doc| + |src| + \sum_j |dup_j|)$.
- Оптимальные параметры алгоритма позволяют поднять PlagDet Score примерно на 15% по сравнению с малоадекватными.