

# Содержание

<b>1. Эксперименты</b>	<b>2</b>
1.1. Методы оценки качества . . . . .	2
1.1.1. Метрика качества $MRR$ . . . . .	2
1.1.2. Метрика качества $TOP_I$ . . . . .	2
1.2. Оптимизация качества WTMF, путём варьирования параметров . . . .	3
1.3. Оптимизация качества WTMF-G, путём варьирования параметров . . .	6
1.4. Сравнительные результаты . . . . .	8

# 1. Эксперименты

Главное целью проведения экспериментов является сравнение двух реализованных методов автоматического установления связей между твитами и новостными статьями: метод основанный на частотности употребления слов и WTMF-G. Для исследования влияния на качество добавления информации о взаимосвязях вида текст-текст также производится сравнительное тестирование методов WTMF и WTMF-G.

Ввиду малого числа твитов в наборах данных тестирование производится на тех же выборках, на которых производится обучение.

## 1.1. Методы оценки качества

Для оценки качества рассматриваются метрики применимые для решения задач информационного поиска. Твит рассматривается как запрос, а список новостей как ответ. Для каждого твита, получаемый список новостей ранжирован по мере убывания их схожести. В работе использованы две метрики:  $MRR$  и  $TOP_I$ , их описание дано ниже.

### 1.1.1. Метрика качества $MRR$

$MRR$  (от англ. Mean reciprocal rank) — статистическая метрика, используемая для измерения качества алгоритмов информационного поиска. Пусть  $rank_i$  — позиция первого правильного ответа в  $i$ -м запросе,  $n$  — общее количество запросов. Тогда значение  $MRR$  можно получить по формуле:

$$MRR = \frac{1}{n} \sum_{i=1}^n \frac{1}{rank_i}.$$

### 1.1.2. Метрика качества $TOP_I$

$TOP_I$  — группа метрик, используемых для оценки качества алгоритмов информационного поиска. Значение метрики  $TOP_I$  численно равно проценту запросов с правильным ответом, входящим в первые  $I$  ответов. Пусть  $n$  — общее количество запросов,  $Q_I(i)$  — равно 1, если правильный ответ на  $i$ -й запрос входит в первые  $I$  предложенных ответов, 0 — в противном случае. Тогда значение  $TOP_I$  можно получить по формуле:

$$TOP_I = \frac{1}{n} \sum_{i=1}^n Q_I(i).$$

В дальнейшем будут рассматриваться следующие три метрики из группы метрик  $TOP_I$ :  $TOP_1$ ,  $TOP_3$ .

## 1.2. Оптимизация качества WTMF, путём варьирования параметров

Оптимизация параметров модели для метода WTMF будет производиться на наборе данных cutted, используя метрику MRR. Модель WTMF зависит от четырёх параметров:  $K$ ,  $I$ ,  $\lambda$ ,  $w_m$ . Параметры  $K$  и  $I$  влияют на время построения модели, а параметры  $\lambda$  и  $w_m$  не влияют на время построения модели.

В качестве начального приближения берутся значения параметров, которое использовали авторы работы [?], а именно:  $K = 30$ ,  $I = 3$ ,  $\lambda = 20$ ,  $w_m = 0.1$ .

Оптимизируются параметры, не влияющие на время работы алгоритма:  $\lambda$  и  $w_m$ . Для этого фиксируются остальные параметры:  $I = 1$ ,  $K = 30$ . Для начала находится оптимальный порядок значений начального приближения. Результаты занесены в таблицу 1.

Таблица 1: Качество работы алгоритма WTMF для различных значений  $\lambda$  и  $w_m$  при фиксированных значениях  $I = 1$ ,  $K = 30$ .

$\lambda \backslash w_m$	<b>0.001</b>	<b>0.01</b>	<b>0.1</b>	<b>1</b>	<b>10</b>	<b>100</b>
<b>0.2</b>	0.6855	0.6877	0.7482	0.3651	0.1526	0.1485
<b>2</b>	0.7000	0.7015	0.7173	0.7525	0.3707	0.1605
<b>20</b>	0.6964	0.7081	0.7149	0.7308	0.7507	0.3784
<b>200</b>	0.7075	0.6991	0.7010	0.7016	0.7146	0.7448
<b>2000</b>	0.6970	0.7070	0.6991	0.7114	0.6994	0.7044

Как видно из таблицы 1 в целом получена достаточно однородная картина для всех порядков  $\lambda$  и  $w_m$ . Заметное снижение качества происходит при большом порядке  $w_m$  и малом порядке  $\lambda$ . Максимальное значение метрики достигнуто при  $\lambda = 2$  и  $w_m = 1$ . Для уточнения значения коэффициентов, производится исследование качества работы алгоритма в окрестностях максимального значения метрики. Результаты приведены в таблице 2.

Из таблицы 2 получаем оптимальные значения коэффициентов  $\lambda = 0.95$  и  $w_m = 1.95$ .

Оптимизируются параметры, влияющие на время работы алгоритма:  $K$  и  $I$ . Для этого фиксируются остальные параметры:  $\lambda = 0.95$ ,  $w_m = 1.95$ . Для начала находится примерное значение коэффициента  $K$  и оптимальное значение  $I$ . Результаты занесены в таблицу 3.

Таблица 2: Качество работы алгоритма WMTF для различных значений  $\lambda$  и  $w_m$  при фиксированных значениях  $I = 1$ ,  $K = 30$ .

$\lambda \backslash w_m$	<b>0.9</b>	<b>0.95</b>	<b>1</b>	<b>1.1</b>	<b>1.2</b>
<b>1.9</b>	0.7442	0.7451	0.7536	0.7542	0.7544
<b>1.95</b>	0.7447	0.7554	0.7452	0.7439	0.7504
<b>2</b>	0.7507	0.7528	0.7504	0.7515	0.7566
<b>2.05</b>	0.7413	0.7505	0.7424	0.7525	0.7479
<b>2.1</b>	0.7405	0.7484	0.7485	0.7502	0.7501

Таблица 3: Качество работы алгоритма WMTF для различных значений  $K$  и  $I$  при фиксированных значениях  $\lambda = 0.95$ ,  $w_m = 1.95$ .

$K \backslash I$	<b>1</b>	<b>2</b>	<b>3</b>
<b>5</b>	0.1232	0.1593	0.1838
<b>10</b>	0.3521	0.4102	0.4437
<b>30</b>	0.7426	0.7422	0.7158
<b>60</b>	0.8326	0.8117	0.7620

Как видно из таблицы 3 увеличение  $K$  приводит к значительному улучшению качества работы алгоритма, увеличении  $I$  приводит к улучшению качества алгоритма только при малых значениях параметра  $K$ , при больших значениях  $K$  увеличение параметра  $I$  приводит к ухудшению качества. Максимальное значение метрики достигнуто при  $K = 60$  и  $I = 1$ . Для уточнения значения коэффициента  $K$ , производится исследование качества работы алгоритма при фиксированном значении коэффициента  $I$ . Результаты приведены в таблице 4.

Из таблицы 4 получаем оптимальные значения коэффициента  $K = 90$

В итоге оптимизации качества рекомендаций на основе алгоритма WMTF были получены оптимальные параметры:  $K = 90$ ,  $I = 1$ ,  $\lambda = 0.95$ ,  $w_m = 1.95$ .

Таблица 4: Качество работы алгоритма WMTF для различных значений  $K$  при фиксированных значениях  $I = 1$ ,  $\lambda = 0.95$ ,  $w_m = 1.95$ .

<b>K</b>	<b>Значение метрики RR</b>
<b>10</b>	0.3595
<b>20</b>	0.6460
<b>30</b>	0.7496
<b>40</b>	0.8003
<b>50</b>	0.8220
<b>60</b>	0.8424
<b>70</b>	0.8472
<b>80</b>	0.8535
<b>82</b>	0.8549
<b>84</b>	0.8597
<b>86</b>	0.8592
<b>88</b>	0.8572
<b>90</b>	0.8675
<b>92</b>	0.8580
<b>94</b>	0.8604
<b>96</b>	0.8612
<b>98</b>	0.8644
<b>100</b>	0.8655
<b>110</b>	0.8627

### 1.3. Оптимизация качества WTMF-G, путём варьирования параметров

Оптимизация параметров модели для метода WTMF-G будет производиться на наборе данных cutted, используя метрику MRR. Модель WTMF-G зависит от пяти параметров:  $K$ ,  $I$ ,  $\lambda$ ,  $\delta$ ,  $w_m$ . Параметры  $K$  и  $I$  влияют на время построения модели, а параметры  $\lambda$  и  $w_m$  не влияют на время построения модели.

В качестве начального приближения параметров взяты оптимальные параметры для метода WTMF, а именно  $K = 90$ ,  $I = 1$ ,  $w_m = 1.95$ ,  $\lambda = 0.95$ . В качестве начального приближения параметра  $\delta$  берем значение 0.1.

Сначала оптимизируем параметры, влияющие на регуляризующий член:  $\lambda$  и  $\delta$ . Для этого фиксируем остальные параметры:  $I = 1$ ,  $K = 90$ ,  $w_m = 1.95$ . Сначала найдём оптимальный порядок значений начального приближения. Результаты занесены в таблицу 9.

Таблица 5: Качество работы алгоритма WMTF-G для различных значений  $\lambda$  и  $\delta$  при фиксированных значениях  $I = 1$ ,  $K = 90$ ,  $w_m = 1.95$ .

$\lambda \backslash \delta$	<b>0.001</b>	<b>0.01</b>	<b>0.1</b>	<b>1</b>	<b>10</b>
<b>0.01</b>	0.3889	0.3842	0.3924	0.3900	0.3895
<b>0.1</b>	0.4895	0.4875	0.4886	0.4850	0.4847
<b>1</b>	0.8227	0.8256	0.8242	0.8225	0.8212
<b>10</b>	0.8477	0.8440	0.8496	0.8454	0.8495
<b>100</b>	0.8294	0.8318	0.8283	0.8240	0.8243

Как видно из таблицы 9 порядок параметра  $\delta$  оказывает влияние на качество, но достаточно слабое, порядок параметра  $\lambda$ , напротив очень сильно влияет на получаемое качество. Максимальное значение метрики достигнуто при  $\lambda = 10$  и  $\delta = 0.1$ . Для уточнения значения коэффициентов, производится исследование качества работы алгоритма в окрестностях максимального значения метрики. Результаты приведены в таблице 10.

Таблица 6: Качество работы алгоритма WMTF-G для различных значений  $\lambda$  и  $\delta$  при фиксированных значениях  $I = 1$ ,  $K = 90$ ,  $w_m = 1.95$ .

$\lambda \backslash \delta$	<b>0.06</b>	<b>0.08</b>	<b>0.1</b>	<b>0.12</b>	<b>0.14</b>
<b>6</b>	0.8589	0.8524	0.8511	0.8580	0.8493
<b>8</b>	0.8483	0.8528	0.8539	0.8439	0.8498
<b>10</b>	0.8504	0.8455	0.8416	0.8453	0.8408
<b>12</b>	0.8453	0.8398	0.8472	0.8376	0.8415
<b>14</b>	0.8462	0.8456	0.8387	0.8398	0.8377

В таблице 10 получена достаточно однородная картина. Возьмём в качестве оптимального значения коэффициент полученную точку максимум:  $\lambda = 6$ ,  $\delta = 0.06$ .

Рассмотрим влияние параметра  $w_m$  и найдём его оптимальное значение. Сначала рассмотрим качество алгоритма для различных порядков  $w_m$ . Результаты занесены в таблицу ??

варируем  $w_m$

Таблица 7: Качество работы алгоритма WMTF-G для различных значений  $w_m$  при фиксированных значениях  $I = 1$ ,  $K = 90$ ,  $\lambda = 6$ ,  $\delta = 0.6$ .

$w_m$	0.01	0.05	0.1	0.5	1	5	10	50	100
<b>MRR</b>	0.8283	0.8296	0.8285	0.8359	0.8442	0.8639	0.8391	0.6094	0.5035

0.01 0.8283 0.05 0.8296 0.1 0.8285 0.5 0.8359 1 0.8442 5 0.8639 10 0.8391 50 0.6094 100 0.5035

1.5 0.8474 2.0 0.8507 2.5 0.8563 3.0 0.8592 3.5 0.8585 4.0 0.8594 4.5 0.8603 5.0 0.8641 5.5 0.8591 6.0 0.8586 6.5 0.8574 7.5 0.8536

берём  $w_m=5$ , варьируем  $K/I$

Оптимизация параметров ещё не завершена, существующая и очень, очень грубая оценка приведена ниже

Оптимизация параметров модели для метода WTMF-G будет производиться на наборе данных `auto_cleared`, используя метрику MRR. Модель WTMF зависит от четырёх параметров:  $K$ ,  $I$ ,  $\delta$ ,  $w_m$ . Параметры  $K$  и  $I$  влияют на время построения модели, а параметры  $\lambda$  и  $w_m$  не влияют на время построения модели.

В качестве начального приближения параметров взяты оптимальные параметры для метода WTMF, а именно  $K = 90$ ,  $I = 1$ ,  $w_m = 1.95$ . В качестве начального приближения параметра  $\delta$  мы берем значение 0.1

Оптимизируется параметр  $\delta$ . Для этого фиксируются остальные параметры:  $K = 90$ ,  $I = 1$ ,  $w_m = 1.95$ . Для начала находится оптимальный порядок значений начального приближения. Результаты занесены в таблицу 9. Как видно из таблицы 9 максимальное значение метрики получено при  $\delta = 0.1$ . Для уточнения значения коэффициента  $\delta$ , производится исследование качества работы алгоритма в окрестностях максимального значения метрики. Результаты приведены в таблице 10. Из таблицы 10 получаем оптимальные значения коэффициента  $\delta = 0.1$ .

В итоге оптимизации качества рекомендаций на основе алгоритма WMTF-G были получены оптимальные параметры:  $K = 90$ ,  $I = 1$ ,  $\delta = 0.1$ ,  $w_m = 1.95$ .

Таблица 8: Качество работы алгоритма WMTF-G для различных значений  $K$  и  $I$  при фиксированных значениях  $w_m = 5$ ,  $\lambda = 6$ ,  $\delta = 0.06$ .

$K \backslash I$	1	2	3	4	5
30	0.7529	0.7927	0.7577	0.6736	0.5794
40	0.7992	0.8194	0.7695	0.6813	0.5828
50	0.8269	0.8349	0.7834	0.6830	0.5801
60	0.8419	0.8450	0.7984	0.7056	0.6006
70	0.8557	0.8466	0.7977	0.7036	0.6002
80	0.8614	0.8511	0.7990	0.7032	0.5957
90	0.8606	0.8522	0.8039	0.7088	0.6038
100	0.8606	0.8527	0.8022	0.7089	0.6021
110	0.8686	0.8553	0.8074	0.7123	0.6065
120	0.8693	0.8579	0.8097	0.7174	0.6085
130	0.8725	0.8588	0.8160	0.7264	0.6206
140	0.8740	0.8597	0.8157	0.7248	0.6241
150	0.8763	0.8620	0.8171	0.7263	0.6200
160	0.8740	0.8596	-	-	-
170	0.8768	0.8606	-	-	-
180	0.8785	0.8613	-	-	-
190	0.8767	0.8616	-	-	-
200	0.8769	0.8613	-	-	-
210	0.8786	0.8613	-	-	-
220	0.8816	0.8632	-	-	-
230	0.8814	0.8646	-	-	-

Таблица 9: Качество работы алгоритма WMTF-G для различных значений  $\delta$  при фиксированных значениях  $K = 90$ ,  $I = 1$ ,  $w_m = 1.95$ .

$\delta$	Значение метрики RR
<b>0.001</b>	0.5508
<b>0.01</b>	0.5307
<b>0.1</b>	0.5695
<b>1</b>	0.5311
<b>10</b>	0.5303
<b>100</b>	0.5203

## 1.4. Сравнительные результаты

Для выявления влияния добавления связей текст-текст на результаты работы метода WMTF-G производится сравнительное тестирование алгоритма WTMF и WTMF-G. Результаты тестирования приведены в таблице 11.

Как видно из таблицы 11 ...

Сравним метод основанный на частотности употребления слов и WTMF-G.



Таблица 10: Качество работы алгоритма WMTF-G для различных значений  $\delta$  при фиксированных значениях  $K = 90$ ,  $I = 1$ ,  $w_m = 1.95$ .

$\delta$	Значение метрики RR
<b>0.05</b>	0.5340
<b>0.1</b>	0.5695
<b>0.15</b>	0.5380
<b>0.25</b>	0.5533
<b>0.3</b>	0.5195
<b>0.35</b>	0.5329

Таблица 11: Сравнительное тестирование алгоритмов WTMF и WTMF-G.

Набор данных	Метрика MRR		Метрика $TOP_1$		Метрика $TOP_3$	
	WTMF	WTMF-G	WTMF	WTMF-G	WTMF	WTMF-G
manual	0.7293	0.	0.	0.	0.	0.
auto	0.8640	0.	0.	0.	0.	0.
total	0.8196	0.	0.	0.	0.	0.
cutted	0.8630	0.	0.	0.	0.	0.
manual_nt	0.6194	0.	0.	0.	0.	0.
auto_nt	0.5297	0.5695	0.	0.	0.	0.
total_nt	0.5729	0.	0.	0.	0.	0.
cutted_nt	0.6495	0.	0.	0.	0.	0.

Метод основанный на частотности употребления слов обозначим как TF-IDF. Результаты тестирования приведены в таблице 12.

Таблица 12: Сравнительное тестирование алгоритмов TF-IDF и WTMF-G.

Набор данных	Метрика MRR		Метрика $TOP_1$		Метрика $TOP_3$	
	TF-IDF	WTMF-G	TF-IDF	WTMF-G	TF-IDF	WTMF-G
manual	0.8336	0.	0.	0.	0.	0.
auto	0.8817	0.	0.	0.	0.	0.
total	0.8610	0.	0.	0.	0.	0.
cutted	0.9075	0.	0.	0.	0.	0.
manual_nt	0.7565	0.	0.	0.	0.	0.
auto_nt	0.6048	0.5695	0.	0.	0.	0.
total_nt	0.6914	0.	0.	0.	0.	0.
cutted_nt	0.7485	0.	0.	0.	0.	0.

Как видно из таблицы 12 ...

объяснение влияния различных датасетов, специфики русского языка и сравнение с результатами статьи.