Выбор факторов для классификации веб-страниц и веб-сайтов

© М.С. Агеев, Б.В. Добров, Н.В. Лукашевич

Научно-исследовательский вычислительный центр
МГУ имени М.В.Ломоносова
АНО Центр информационных исследований
{ageev, dobroff, louk}@mail.cir.ru

Аннотация

В статье описываются подходы, использованные коллективом разработчиков Университетской информационной системы РОССИЯ (УИС РОССИЯ, http://uisrussia.msu.ru), для выполнения заданий РОМИП 2010 по классификации веб-страниц и веб-сайтов.

1. Введение

В цикле РОМИП 2010 мы приняли участие в дорожках классификации веб-страниц и веб-сайтов.

Опыт участия в данной дорожке [1] показал, что задача классификации сайтов имеет следующие особенности:

- обучающая выборка – сайты из каталога DMOZ – имеет определенную структуру, которую необходимо учитывать:
- не все страницы сайтов обучающей выборки являются релевантными рубрике (но некоторые страницы все же должны быть релевантны) [3];
- в то же время, имеет место неполнота описаний сайтов обучающей выборки рубриками, так как сайты отнесены не более чем к одной рубрике.

2. Классификация веб-страниц

В этом году мы исследовали следующие подходы к классификации веб-страниц:

1) Методы машинного обучения в форме линейного (a-la Байесовского) классификатора документов, в котором векторное представление рубрики строится на основе учета распределения слова по документам, сайтам и рубрикам. Предполагается, что «хорошее» слово должно
   - присутствовать на различных сайтах, релевантных данной рубрике, но редко на других сайтах коллекции;
   - присутствовать в нескольких документах, релевантных сайтов;
   - «хороших» слов не может быть много [7].
2) Быстрое экспертное описание смысла рубрик на основе Общественно-политического тезауруса [4, 5, 6].

2.1. Классификация на основе машинного обучения

Выполнение задания разбито на четыре этапа:

1) построение векторного представления документов обучающей коллекции DMOZ;
2) построение векторного представления рубрик;
3) классификация документов BY.WEB путем вычисления наиболее близких векторов рубрик по «косинусной» мере;
4) выбор наиболее релевантных документов для рубрики, с учетом различных порогов, влияющих на соотношение полноты/точности.

На первом этапе производится приведение всех слов к нормальной форме (морфологический анализ) и вычисление весов слов по модифицированной формуле TF*IDF BM25 [2]. Выбирались только русскоязычные слова (слова, состоящие из русских символов).

При анализе результатов мы обнаружили, что значительная часть коллекции DMOZ (12% документов) имеет кодировку KOI8. К сожалению, мы не провели предварительное преобразование кодировки обучающей коллекции.

На втором этапе для каждой пары рубрика-слово производится вычисление следующих факторов:

- coll_feat_docs — количество документов, содержащих слово, в коллекции;
- coll_feat_sites — количество сайтов, содержащих слово, в коллекции;
- coll_feat_cats — количество категорий, содержащих слово, в коллекции;
- cat_docs — количество документов для данной категории;
- cat_sites — количество сайтов для данной категории;
- cat_feat_docs — количество документов для данной категории, содержащих данное слово
- cat_feat_sites — количество сайтов для данной категории, содержащих слово;
- sites_tf=cat_feat_sites/cat_sites — доля сайтов для категории, содержащих данное слово
- sum_sites_tf — сумма sites_tf по всем категориям, содержащим данное слово
- weight — вес данного слова для категории по формуле

$$\text{weight} = \text{sites\_tf}\,\frac{\text{sites\_tf}}{\text{sum\_sites\_tf}} \qquad (1)$$

В качестве векторного представления рубрики выбирается не более 100 слов с наибольшим весом, для которых выполняются два условия:

- cat_feat_docs≥2 — слово содержится не менее, чем в двух релевантных документах;
- cat_feat_docs/cat_docs>0.05 — доля документов категории, содержащих данное слово, более 5%.

Указанные пороги на количество слов и документов подбирались на основе выборочного ручного анализа списка слов в получаемом векторном представлении рубрик. Выборочный анализ показал, что слова вполне соответствуют смыслу рубрики, наверх списка «всплывают» наиболее важные слова, к концу списка появляется множество слаборелевантных слов.

На третьем этапе производится классификация документов BY.WEB путем вычисления наиболее близких векторов рубрик. Мы использовали два варианта меры близости между векторами:

а) «косинусная» мера близости:

$$\cos(d,r) = \frac{\sum_i d_i \cdot r_i}{\sqrt{\sum_i d_i^2}\sqrt{\sum_i r_i^2}} \quad (2)$$

где d_i — веса нормализованных слов документа, вычисленных по модифицированной формуле TF*IDF BM25 [2], r_i — веса слов векторного представления рубрики, вычисленные по формуле (1).

b) скалярное произведение нормализованного вектора документа на вектор рубрики:

$$\mathrm{dist}(d,r) = \frac{\sum_i d_i \cdot r_i}{\sqrt{\sum_i d_i^2}}$$

Предположительно, вариант «b» должен давать более низкий вес рубрикам, для которых векторное представление построилось плохо — количество слов в векторном представлении менее 100 и/или слова имеют низкий вес, то есть редко встречаются в на сайтах категории.

Для каждого документа вычисляются 10 наиболее близких категорий.

Итак, на третьем этапе для каждой категории получен список из ровно 10 наиболее близких категорий. Для формирования списка документов, отнесенных к рубрике, следует ограничить количество выбираемых документов с целью выбора оптимального соотношения полноты/точности. Так как априори количество релевантных документов для категорий неизвестно, мы создали 4 прогона с различными порогами отсечения документов:

Прогон «**vn1**» (больше точность): используется косинусная мера близости; для каждого документа выдается не более 5 рубрик, находящихся на расстоянии не менее **0.02**, но при этом количество документов для рубрики должно быть не менее 50 (иначе ограничения смягчаются); для рубрики выдается не более 5000 документов.

Прогон «**vn2**» (больше полнота): используется косинусная мера близости; для каждого документа выдается не более 5 рубрик, находящихся на расстоянии не менее **0.005**, но при этом количество документов для рубрики должно быть не менее 50 (иначе

ограничения смягчаются); для рубрики выдается не более 10000 документов.

Прогон «**vn3**» (среднее значение полноты/точности): используется косинусная мера близости; для каждого документа выдается не более 5 рубрик, находящихся на расстоянии не менее **0.01**, но при этом количество документов для рубрики должно быть не менее 50 (иначе ограничения смягчаются); для рубрики выдается не более 1000 документов.

Прогон «**vu**» (ненормализованные вектора рубрик): используется скалярное произведение; для каждого документа выдается не более 5 рубрик, находящихся на расстоянии не менее **0.02**, но при этом количество документов для рубрики должно быть не менее 50 (иначе ограничения смягчаются); для рубрики выдается не более 5000 документов.

2.2. Классификация на основе экспертного описания смысла рубрик

Мы создали 3 прогона с различными порогами отсечения документов:

Прогон «**a21**» (без ограничений): выдаются все рубрики документа, имеющие вес не менее 60%.

Прогон «**a22**» (меньше рубрик на документ): для каждого документа выдается не более 5 рубрик, имеющих вес не менее 60%.

Прогон «**a23**» (выше точность): для каждого документа выдается не более 5 рубрик, имеющих вес не менее 70%.

Прогон «**e4**»: используется описания рубрик и программа обработки АЛОТ в варианте, который мы использовали для этой же дорожки в 2007 году [1]. К сожалению, при обработке была допущена ошибка, поэтому корректно оценить качество работы алгоритма не представляется возможным.

Еще 3 прогона являются вариацией прогонов a21-a23, но содержали ошибку, их мы назовем «**e1**», «**e2**», «**e3**» соответственно.

2.3. Результаты классификации веб-страниц

В таблице 1 приведены результаты всех участников дорожки классификации веб-страниц по таблице релевантности AND_JUDGEDONLY (учитываются только оцененные документы, релевантным считается документ, который **все** асессоры оценили как релевантный). Прогоны отсортированы по убыванию F-меры.

В таблице 2 приведены результаты по таблице релевантности OR_JUDGEDONLY (учитываются только оцененные документы, релевантным считается документ, который **хотя бы один** асессор оценил как релевантный). Прогоны отсортированы по убыванию F-меры.

RUN	F1	F1 (micro average)	Precision	Precision (micro average)	Recall	Recall (micro average)	Accuracy
vn2	**55%**	**59%**	65%	68%	**52%**	**52%**	**98%**
xxxx-3	53%	58%	64%	66%	50%	51%	98%
a1	45%	48%	55%	54%	45%	43%	98%
a2	45%	47%	55%	54%	44%	42%	98%
a3	43%	45%	55%	53%	42%	40%	98%
e1	41%	43%	55%	52%	39%	36%	98%
e2	41%	42%	54%	52%	39%	35%	98%
vn3	40%	41%	**73%**	74%	30%	28%	98%
vn1	40%	45%	72%	**76%**	30%	31%	98%
e3	39%	40%	54%	51%	36%	32%	98%
vu	38%	42%	64%	69%	29%	30%	98%
xxxx-1	36%	42%	49%	60%	32%	32%	98%
xxxx-2	26%	30%	40%	52%	21%	21%	98%
e4	16%	16%	36%	37%	11%	10%	97%

лучший результат
95% и более от лучшего
90% и более от лучшего

Таблица 1. Результаты классификации веб-страниц, AND_JUDGEDONLY

RUN	F1	F1 (micro average)	Precision	Precision (micro average)	Recall	Recall (micro average)	Accuracy
xxxx-3	**54%**	58%	83%	86%	**45%**	**44%**	98%
vn2	53%	**58%**	81%	88%	42%	43%	**98%**
a21	52%	54%	80%	80%	43%	41%	97%
a22	51%	54%	80%	80%	43%	40%	97%
a23	49%	52%	80%	79%	41%	39%	97%
e1	48%	50%	81%	80%	39%	36%	97%
e2	47%	49%	81%	80%	38%	36%	97%
e3	45%	47%	80%	79%	36%	34%	97%
xxxx-1	38%	41%	68%	76%	30%	28%	97%
vn1	34%	37%	84%	**89%**	23%	24%	97%
vu	33%	37%	78%	85%	23%	23%	97%
vn3	33%	35%	**84%**	87%	22%	22%	97%
xxxx-2	26%	28%	55%	66%	19%	18%	97%
e4	19%	20%	62%	65%	12%	12%	97%

Таблица 2. Результаты классификации веб-страниц, OR_JUDGEDONLY

Результаты показывают, что прогон «**vn2**» (метод машинного обучения, более ориентированный на полноту) показал лучший результат по F-мере при использовании «строгой» релевантности, и один из двух лучших результатов при использовании «слабой» таблицы релевантности.

3. Классификация веб-сайтов

Классификация веб-сайтов коллекции BY.WEB осуществлялась на основе классификации отдельных страниц сайтов. Для определения релевантности сайта учитывалось количество релевантных страниц сайта, вес рубрики страницы, флаг «является ли страница главной на сайте», а также определенные пороги, регулирующие соотношение полноты/качества классификации сайтов.

Для каждой пары сайт-рубрика следующие факторы вычисляются на основе весов рубрики для документов сайта:

- sum_rubr_weight — суммарный вес данной рубрики по всем документам сайта;
- cnt_rubr — количество документов сайта, отнесенных к данной рубрике (с ненулевым весом);
- cnt_rubr(threshold) — количество документов сайта, отнесенных к рубрике с весом не менее threshold;
- cnt_pages — общее количество документов сайта;
- mp_weight — вес главной страницы сайта

Вес пары сайт-рубрика вычислялся для различных прогонов по-разному, использовались следующие способы вычисления:

a) для классификации на основе методов машинного обучения:

$$W(s,r) = \log_{10}\left(\begin{array}{l} \dfrac{\text{mp\_weight}+0.01}{1+0.01}\cdot\min\left(\text{mp\_weight},0.1\right) \\ \cdot\dfrac{\log_2\left(\text{cnt\_rubr}\left(0.1\right)+2\right)}{\log_2\left(\text{cnt\_pages}+2\right)}\cdot\min\left(\dfrac{\text{cnt\_rubr}}{\text{cnt\_pages}},0.1\right) \\ \cdot\dfrac{\log_2\left(\text{sum\_rubr\_weight}+2\right)}{\log_2\left(\text{cnt\_pages}+2\right)} \end{array}\right)$$

b) для классификации на основе экспертного описания рубрик, первый вариант:

$$W(s,r)=\left(\begin{array}{l}\dfrac{\text{mp\_weight}+0.1}{1+0.1}\cdot\left(\text{mp\_weight}>0.8\,?\,1:0.4\right)\\ \cdot\dfrac{\log_2\left(\text{cnt\_rubr}+1\right)}{\log_2\left(\text{cnt\_pages}+1\right)}\cdot\dfrac{\log_2\left(\text{cnt\_rubr}\left(0.8\right)+2\right)}{\log_2\left(\text{cnt\_pages}+2\right)}\\ \cdot\min\left(\dfrac{\text{cnt\_rubr}}{\text{cnt\_pages}},0.1\right)\end{array}\right)$$

с) для классификации на основе экспертного описания рубрик, второй вариант (формула 2007-го года [1]):

$$W(s,r)=\left(\begin{array}{l}\dfrac{\text{sum\_rubr\_weight}}{\text{cnt\_rubr}}\cdot\dfrac{\text{cnt\_rubr}\left(0.8\right)+1}{\text{cnt\_pages}+1}\\ \cdot\max\left(\dfrac{\text{cnt\_rubr}}{200},1\right)\cdot\max\left(\dfrac{\text{cnt\_rubr}}{0.4\cdot\text{cnt\_pages}},1\right)\end{array}\right)$$

На основе комбинирования различных формул взвешивания сайтов, методов рубрикации и различных порогов отсечения документов, мы создали 8 прогонов:

Прогон «**vn10**» (машинное обучение, максимальная полнота): используется классификация документов на основе машинного обучения с косинусной мерой близости, вес пары сайт-рубрика определяется по формуле «а», для каждого сайта выдается не более 5 рубрик, для каждой рубрики выдается не менее 5, но и не более 100 сайтов, без ограничения по весу.

Прогон «**vn5.5**» (машинное обучение, повышение точности): используется алгоритм прогона «vn10», но из выдачи исключаются сайты с весом менее (–5.5) (вес сайтов отрицательный из-за использования логарифма).

Прогон «**vn5**» (машинное обучение, повышение точности): используется алгоритм прогона «vn10», но из выдачи исключаются сайты с весом менее (–5).

Прогон «**vn4**» (машинное обучение, максимальная точность): используется алгоритм прогона «vn10», но из выдачи исключаются сайты с весом менее (–4).

Прогон «**ac**» (АЛОТ, формула 2007-го года): используется классификация документов на основе экспертного описания рубрик, вес пары сайт-рубрика определяется по формуле «с», для каждого сайта выдается не более 5 рубрик, без ограничения по весу.

Прогон «**ac_l100**» (АЛОТ, формула 2007-го года, выше точность): используется алгоритм прогона «ac», плюс

дополнительное ограничение — для каждой рубрики выдается не более 100 сайтов.

Прогон «**ab_l100**» (АЛОТ, формула «b»): используется классификация документов на основе экспертного описания рубрик, вес пары сайт-рубрика определяется по формуле «b», для каждого сайта выдается не более 5 рубрик, для каждой рубрики выдается не более 100 сайтов.

Прогон «**2007**»: используется описания рубрик, программа обработки АЛОТ и формула вычисления весов сайта в варианте, который мы использовали для этой же дорожки в 2007 году [1].

2.3. Результаты классификации веб-сайтов

В таблице 3 приведены результаты всех участников дорожки классификации веб-сайтов по таблице релевантности OR_JUDGEDONLY (учитываются только оцененные сайты, релевантным считается сайт, который **все** асессоры оценили как релевантный). Прогоны отсортированы по убыванию F-меры.

В таблице 4 приведены результаты по таблице релевантности AND_JUDGEDONLY (учитываются только оцененные сайты, релевантным считается сайт, который **хотя бы один** асессор оценил как релевантный). Прогоны отсортированы по убыванию F-меры.

Из таблиц видно, что результаты существенно зависят от таблицы релевантности. Для «слабых» требований релевантности (таблица «OR») метод, основанный на экспертном описании рубрик, показывает лучший результат по F-мере. Метод машинного обучения, ориентированный на максимальную полноту – второй по значению F-меры результат. Как и ожидалось, эти алгоритмы показали высокую полноту классификации.

При рассмотрении «сильных» требований релевантности (таблица «AND») эти методы также показывают наилучшую полноту классификации среди всех прогонов, но проигрывают методам, которые используют более оптимальное соотношение полноты/точности. Это можно объяснить тем фактом, что при «сильных» требованиях релевантности точность классификации более важна, чем при «слабых» требованиях релевантности.

RUN	F1	F1 (micro average)	Precision	Precision (micro average)	Recall	Recall (micro average)	Accuracy
2007	**53%**	**56%**	55%	55%	**63%**	**58%**	98%
ac	49%	54%	53%	54%	55%	54%	98%
xxxx-9	48%	54%	59%	65%	47%	46%	98%
xxxx-7	47%	51%	65%	74%	40%	39%	98%
xxxx-6	46%	53%	64%	73%	40%	42%	**98%**
ac_l100	40%	38%	63%	66%	34%	27%	98%
xxxx-4	39%	47%	56%	75%	35%	34%	98%
xxxx-2	36%	42%	67%	75%	29%	30%	98%
ab_l100	36%	35%	58%	59%	31%	25%	98%
xxxx-5	34%	37%	55%	78%	28%	25%	98%
vn10	27%	31%	55%	60%	22%	21%	98%
vn5.5	26%	30%	55%	60%	21%	20%	98%
xxxx-3	23%	22%	**71%**	**88%**	15%	12%	98%
xxxx-1	22%	18%	61%	80%	16%	10%	98%
vn5	22%	24%	58%	59%	16%	15%	98%
xxxx-8	20%	22%	61%	79%	14%	13%	98%
vn4	16%	17%	57%	52%	10%	10%	97%

лучший результат

95% и более от лучшего

90% и более от лучшего

Таблица 3. Результаты классификации веб-сайтов, OR_JUDGEDONLY

RUN	F1	F1 (micro average)	Precision	Precision (micro average)	Recall	Recall (micro average)	Accuracy
xxxx-7	**48%**	**52%**	46%	50%	58%	55%	99%
xxxx-6	43%	48%	42%	44%	55%	53%	98%
2007	41%	44%	34%	32%	**73%**	**72%**	98%
xxxx-9	40%	43%	38%	36%	55%	53%	98%
ac	40%	45%	33%	33%	65%	70%	98%
xxxx-2	39%	46%	**50%**	51%	41%	42%	99%
ac_l100	39%	40%	42%	44%	46%	37%	98%
xxxx-5	37%	42%	37%	52%	43%	35%	**99%**
xxxx-4	36%	42%	36%	43%	50%	41%	98%
ab_l100	34%	37%	37%	39%	41%	35%	98%
vn10	31%	33%	37%	39%	33%	29%	98%
vn5.5	30%	32%	39%	39%	31%	27%	98%
vn5	28%	27%	44%	38%	25%	21%	98%
xxxx-3	27%	26%	47%	56%	25%	17%	99%
xxxx-8	23%	26%	45%	50%	20%	18%	99%
xxxx-1	21%	23%	41%	**56%**	17%	15%	99%
vn4	21%	19%	42%	32%	16%	13%	98%

лучший результат

95% и более от лучшего

90% и более от лучшего

Таблица 4. Результаты классификации веб-сайтов, AND_JUDGEDONLY

Заключение

Наш коллектив использует возможность участия в РОМИП как эффективный способ уточнения параметров применяемых нами методов.

В этом году мы исследовали методы классификации веб-страниц и веб-сайтов:

- методы машинного обучения, которые учитывают определенную специфику обучения на выборке сайтов, представленных в веб-каталоге;
- методы классификации, основанные на экспертном описании рубрик, не зависящем от обучающей выборки.

Представленные методы также учитывают следующую специфику рассматриваемой задачи: обучающая выборка не позволяет оценить количество документов (сайтов), релевантных той или иной рубрике в коллекции тестирования. Поэтому для каждого алгоритма представлено несколько вариаций, ориентированных на различное соотношение полноты и точности классификации.

На обеих дорожках классификации лучшие (по F-мере) результаты показали варианты алгоритмов, нацеленных в большей степени на полноту классификации.

Литература

1. Агеев М.С., Добров Б.В., Красильников П.В., Лукашевич Н.В., Павлов А.М., Сидоров А.В., Штернов С.В. УИС РОССИЯ в РОМИП'2007: поиск и классификация // Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2007-2008: Семинар в рамках Всероссийской науч. конф. RCDL'2007. 18 окт. 2007 г., Переславль-Залесский - изд-во Санкт-Петербург: НУ ЦСИ, 2008, 258 с. - С.199-220. http://www.cir.ru/docs/ips/publications/2007_romip_uis.pdf
2. Агеев М.С., Добров Б.В., Лукашевич Н.В., Сидоров А.В. Экспериментальные алгоритмы поиска/классификации и сравнение с "basic line" // Российский семинар по Оценке Методов Информационного Поиска (РОМИП 2004): Семинар в рамках Всероссийской науч. конф. RCDL'2004. 1 окт. 2004 г. - Пущино, 2004. - С.62-89. http://www.cir.ru/docs/ips/publications/2004_romip_uis.pdf

3. Васильев В.Г. Обработка и классификация документов с использованием системы СКАТ. // Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2009: Семинар в рамках Всероссийской науч. конф. RCDL'2009. - изд-во Санкт-Петербург: НУ ЦСИ, 2009 - С.141-150.
http://romip.ru/romip2009/13_skat.pdf
4. Лукашевич Н.В., Автоматическое рубрицирование потоков текстов по общественно-политической тематике // НТИ. Сер.2., 1996. - № 10. - С.22-30.
5. Лукашевич Н.В. Автоматизированное формирование информационно-поискового тезауруса по современной общественно-политической жизни России // НТИ. Сер.2. 1995. No 3. С.22-24.
6. Лукашевич Н.В., Добров Б.В., Тезаурус русского языка для автоматической обработки больших текстовых коллекций // Компьютерная лингвистика и интеллектуальные технологии. Труды Международного семинара Диалог'2002 / Под ред. А.С.Нариньяни - М.: Наука - 2002. - Т.2 - С.338-346.
http://www.cir.ru/docs/ips/publications/2002_dialog_ruthes.pdf
7. Поляков П.Ю., Плешко В.В., Ермаков А.Е. RCO на РОМИП 2009 // Российский семинар по Оценке Методов Информационного Поиска. Труды РОМИП 2009: Семинар в рамках Всероссийской науч. конф. RCDL'2009. - изд-во Санкт-Петербург: НУ ЦСИ, 2009 - С.122-134
http://romip.ru/romip2009/11_rco.pdf

Feature Selection for Categorization of Web Pages and Web Sites

Mikhail S. Ageev, Boris V. Dobrov, Natalia V. Loukachevitch

In the paper we describe methods used by the team of UIS RUSSIA (http://www.cir.ru/eng/) search engine for ROMIP 2010 tracks. We participated in the web-page classification track and web-sites categorization track. We used machine learning and knowledge-based algorithms, that acknowledge specific structure of training ant test sets in web categorization tasks.