PRIKASPIYSKIY ZHURNAL: Upravlenie i Vysokie Tekhnologii
(CASPIAN JOURNAL: Management and High Technologies), 2014, 1 (25)
PROBLEMS OF ROBOTICS, INSTRUMENTATION,
CREATION AND USE OF NEW MATERIALS

УДК 004.891.2

ПРЕДСТАВЛЕНИЕ ДОКУМЕНТА В ВИДЕ ВЕКТОРА КЛЮЧЕВЫХ ФРАЗ ДЛЯ РЕШЕНИЯ ЗАДАЧИ ПОИСКА ПО УРОВНЮ ТЕХНИКИ В ОПИСАНИЯХ ПАТЕНТОВ

Дыков Михаил Александрович, аспирант, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. им. Ленина, 28, e-mail: dmawork@mail.ru

Кравец Алла Григорьевна, доктор технических наук, профессор, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. им. Ленина, 28, e-mail: agk@gde.ru

Коробкин Дмитрий Михайлович, кандидат технических наук, доцент, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. им. Ленина, 28, e-mail: dkorobkin80@mail.ru

Укустов Сергей Сергеевич, аспирант, Волгоградский государственный технический университет, 400005, Российская Федерация, г. Волгоград, пр. им. Ленина, 28, e-mail: sergey@ukstv.me

Стрелков Олег Игоревич, директор, Федеральный институт промышленной собственности, 123995, Российская Федерация, г. Москва, Бережковская наб., 30, к. 1, e-mail: fips@rupto.ru

Работа посвящена автоматизации решения задачи анализа текстовых документов в отношении уровня описываемой в них техники. Обоснован метод представления патентных документов в виде векторов ключевых фраз, а также метод использования этих векторов в задаче поиска по уровню техники. Разработанные методы призваны значительно уменьшить время, затрачиваемое экспертом на проведение патентной экспертизы. Предложенная для этой цели методика включает в себя последовательное решение нескольких задач: предобработка патентных документов; выделение ключевых фраз из текстов патентных документов; определение меры сходства между векторами ключевых фраз для сравниваемой пары описаний патентов. Достоинством методики является ее масштабируемость на все множество описаний патентов, включающее в себя десятки миллионов документов. Результаты выполненных экспериментов по поиску патентов прототипов, проведенные на выборке из выданных в 2012 г. российских патентов, продемонстрировали значительное превосходство разработанной методики в отношении показателей точности по сравнению с существующими.

Ключевые слова: поиск по уровню техники, патентная экспертиза, морфологический анализ, обработка естественного языка, данные большого объема, патентная экспертиза, поиск ключевых фраз, степени сходства текстовых документов, методика оценки

DOCUMENT REPRESENTATION AS A KEY PHRASES VECTOR FOR PATENTS PRIOR-ART RETRIEVAL

Dykov Mikhail A., post-graduate student, Volgograd State Technical University, 28 Lenin av., Volgograd, 400005, Russian Federation, e-mail: dmawork@mail.ru

Kravets Alla G., D.Sc. (Engineering), Professor, Volgograd State Technical University, 28 Lenin av., Volgograd, 400005, Russian Federation, e-mail: agk@gde.ru

Korobkin Dmitriy M., Ph.D. (Technical), Associate Professor, Volgograd State Technical University, 28 Lenin av., Volgograd, 400005, Russian Federation, e-mail: dkorobkin80@mail.ru

Ukustov Sergey Sergeyevich, post-graduate student, Volgograd State Technical University, 28 Lenin av., Volgograd, 400005, Russian Federation, e-mail: sergey@ukstv.me

Strelkov Oleg I., Director, Federal Institute of Industrial Property, 1 Building, 30 Berezhkovskaya Naberezhnaya, Moscow, 123995, Russian Federation, e-mail: fips@rupto.ru

We proposed a method of patent document representation as a key phrases vector and a method of using these vectors for the patents prior-art retrieval task. These methods are developed to significantly decrease of the time that examiner has to spend during prior-art retrieval. Proposed method includes a solution of a step by step subtasks set: the patents' documents preprocessing, the key phrases retrieval from texts of patents' documents, the similarity calculation between vectors of patents' documents key phrases. The one of the main advantages of proposed methods is their easy scalability for the complete patents set which includes millions of documents. Performed experiments showed that developed methods significantly outperform the baseline.

Keywords: prior-art patent search, patent examination, morphological analysis, natural language processing, big data, patent examination, key phrases search, texts similarity calculation, estimation procedure

С каждым годом все большее количество компаний изъявляют желание запатентовать свои изобретения [3, 4]. Поэтому растет поток заявок на выдачу патентов. По данным Всемирной организации интеллектуальной собственности, в 2012 г. в общей сложности на патентную экспертизу поступило более 2300 тыс. заявок, что более чем на 9 % выше, чем за предыдущий год. Огромный семидесятимиллионный фонд существующих патентов и возрастающий поток заявок увеличивают время их рассмотрения экспертами, поэтому срок экспертизы заявок в различных патентных ведомствах по всему миру, в том числе и в Федеральном институте промышленной собственности (ФИПС), достигает иногда нескольких лет. В ходе экспертизы заявочных материалов эксперт вынужден составлять сотни поисковых запросов и просматривать тысячи существующих патентов. Данная процедура порой занимает десятки часов [12]. Возрастающая нагрузка на патентные офисы привела к необходимости разработки автоматизированных систем информационной поддержки принятия решений экспертами.

Задачу патентной экспертизы можно разбить на две подзадачи:

1) поиск релевантных патентов – поиск по уровню техники;

2) принятие решения о том, опровергает ли релевантный патент новизну заявки или нет.

В данной статье рассматривается автоматизация решения первой задачи.

Многие ученные занимаются вопросом автоматизации поиска по уровню техники. Ежегодно проводится конференция CLEF [7]. В ее рамках осуществляются соревнования среди различных систем по решению задач автоматизации патентной экспертизы, в том числе автоматизации поиска по уровню техники. Различными учеными, а также участниками соревнований в рамках данной конференции были предложены методы, основанные на машинном обучении [13], анализе синтаксических отношений [1], графов цитирования патентов прототипов и классов патентов [11], формирование поискового запроса из заявки и использовании функции ранжирования BM25 [8]. Так, были использованы методы реферирования и аннотирования для составления поискового запроса из текста патента [6], применены внешние базы знаний [2], использованы униграммы и биграммы [5]. Однако предложенные решения не показывают значительного улучшения значения recall по сравнению с базовым методом, основанном на мере TF-IDF [7].

Целью данной работы является создание метода автоматизированного поиска по уровню техники, который бы учитывал недостатки существующих и позволил бы превзойти их показатели точности.

Общая характеристика методики работы с текстами. На первом этапе производится предобработка существующих баз патентов: приведение их к единому формату, морфологический анализ, удаление стоп-слов и слов, относящихся к не основным частям речи. На втором этапе производится выделение множества фраз из текстов патентов и построение на их основе статистических портретов патентов, которые представляют собой векторы от-

носительных частот для фраз. Для каждой поступившей заявки производится предобработка, аналогичная предобработке коллекции существующих патентов. Далее производится построение вектора фраз на основании текста заявки. В заключение производится расчет сходства между текстом заявки и каждым из существующих в базе патентов. На основании результатов расчетов выполняется ранжирование существующих патентов по степени их сходства с текстом поданной заявки.

Предобработка существующих массивов патентов. В ходе предобработки существующих массивов для дальнейшей обработки были выделены следующие поля каждого документа: номер патента; классы патента; дата публикации патента; список цитированных патентов; название; аннотация; описание и формула. Имеющиеся базы патентов США и России были первоначально представлены в трех различных форматах. В ходе работы авторов все они были приведены к единообразному формату.

При предобработке из текстов патентов были исключены 100 наиболее часто встречающихся слов. Далее был выполнен морфологический анализ текстов патентов. Для этой цели были выбраны существующие программные решения: "Stanford POS Tagger" для английского языка и "MyStem" – для русского [9, 10]. Данные инструментальные средства являются свободным программным обеспечением. Их отличительной особенностью является то, что помимо высокой точности определения базовых форм и частей речи у известных слов, они также с большой точностью определяют эти составляющие у неизвестных слов, что является очень важным ввиду наличия множества специфических терминов в текстах патентных документов. Для дальнейшей обработки были выбраны только те слова, которые относятся к следующим частям речи: существительное, прилагательное, глагол, числительное, наречие. Такой выбор определяется тем, что, как правило, именно они несут в себе всю смысловую нагрузку текста.

Извлечение ключевых фраз и поиск релевантных патентов. Под фразой будем понимать набор слов, ограниченный знаком препинания «точка». При этом в контексте описания патента фразы могу быть как неосмысленными, так и осмысленными: понятия предметной области, характеристики объектов, действия над объектами. Целью разработанного метода является выделение осмысленных фраз. Существует ряд подходов по выделению фраз из текстов на естественном языке, в том числе основанных на n-граммах; на выделении семантических и синтаксических зависимостей между словами в предложениях. Недостатком подхода, основанного на выделении n-грамм, является то, что одно и то же понятие в разных местах может быть представлено различными словами в разном порядке с различными промежуточными словами. Недостатки подхода, основанного на выделении синтаксических и семантических зависимостей: низкая скорость синтаксического и семантического разбора; невысокая точность, связанная с появлением ошибок 1-го и 2-го рода. Недостаток точности особенно сильно проявляется применительно к патентным документам, так как в них используются длинные предложения, сложные языковые конструкции и специфические термины. Поэтому с учетом приведенных выше недостатков было решено разработать собственный метод, который был бы не восприимчив к порядку слов и промежуточным словам. Представим документ патента в виде множества предложений:

$$D = (S_1, S_2, S_3 \ldots S_k),$$

где D – документ; S_i – i предложение; k – количество предложений.

$$S = (w_1, w_2, w_3 \ldots w_l),$$

где, w_i – i слово; l – количество слов.

Фразы ищутся среди последовательностей слов:

$$PS = (w_i, w_{i+1}, w_{i+2} \ldots w_{i+n-1}),$$

где n – максимальная длина последовательности; при этом $n \leq l$.

На последовательность PS накладывается ограничение вхождения только в одно предложение. Параметр n не может быть слишком большим, так как чем дальше отстоят друг от друга слова в предложении, тем меньше вероятность того, что они могут образовать осмысленную фразу. Также параметр n не может быть слишком маленьким, так как некоторые осмысленные фразы могут образовываться словами, стоящими не рядом. Опытным путем было установлено рациональное значение параметра $n = 10$.

Из каждой последовательности слов производится выделение фраз:

$$P_o = (w_{i1}, w_{i2}, w_{i3} \ldots, w_{im}),$$

где P_o – o-вая фраза в предложении; w_{ij} – слово из последовательности с произвольным индексом ij; m – максимальная длина фразы.

При этом порядок слов во фразе не имеет значения. Эмпирическим путем было установлено, что рациональное максимальное значение параметра $m = 3$. Всего в каждом предложении имеется $l - n + 1$ последовательностей слов. Каждая фраза из последовательности входящая в множество $(P_2, P_3 \ldots P_{l-n+1})$ содержит последнее слово последовательности.

Таким образом, общее количество фраз в патентном документе рассчитывается по следующей формуле:

$$Total = \sum_{i=1}^{k} \left(\prod_{j=n_i-m}^{n_i-1} j + \sum_{p=2}^{l_i-n_i+1} \prod_{j=n_i-m+1}^{n_i-1} j \right),$$

где n_i – максимальная длина последовательности в i предложении; l_i – количество слов в i предложении.

Патентный документ можно представить в виде вектора фраз:

$$DV = \{(P_1, F_1), (P_2, F_2) \ldots, (P_x, F_x)\},$$

$$F_i = \max_j \frac{Count(P_i, D)}{Count(w_j, Collection)},$$

$$\sum_{i=1}^{x} Count(P_i, D) = Total,$$

$$Count(P_i, D) \geq 2,$$

где F_i – относительная частота i фразы; $Count(P_i, D)$ – количество упоминаний фразы P_i в документе D; $Count(w_j, Collection)$ – количество упоминаний слова w_j во всем массиве патентов $Collection$.

Одной из ключевых особенностей представления документа в виде вектора фраз является то, что в данный вектор входят только те фразы, которые встречаются в документе как минимум два раза. Данное ограничение введено на основании предположения о том, что автор заявки на патент старается несколько раз закрепить основные положения в различных частях патента: в аннотации, в реферате и в формуле. Такое ограничение в комбинации с использованием относительной частоты использования позволило придать максимальный вес ключевым фразам.

Степень сходства между текстами описаний двух патентов определяется по следующей формуле:

$$Similarity(Pat_i, Pat_j) = \sum_{l=1}^{t} F_l^i,$$

$$P_l^i \in \{P_1^i, P_2^i \ldots P_{k1}^i\} \cap \{P_1^j, P_2^j \ldots P_{k2}^j\}.$$

При поиске по уровню техники учитывается только значимость фраз в заявке. При этом не учитывается, насколько значима фраза в сравниваемом патенте, так как бывают случаи, когда в целом ключевая особенность патента отлична от заявки, но некоторые его положения релевантны тематике рассматриваемой заявки.

Таким образом, для нахождения патентов, релевантных рассматриваемой заявке, производится расчет ее схожести с каждым патентом из имеющейся базы. Далее производится ранжирование заявок в базе на основании рассчитанной схожести.

Результаты тестовых экспериментов. Для тестирования разработанных методов в качестве экспериментального материала были взяты 200 существующих патентов класса H01 из российской базы за 2012 г., которые выступали в эксперименте в качестве тестовых заявок: для них нужно было произвести поиск по уровню техники. Данная выборка осуществлялась случайным образом из полной выборки, состоящей из 1306 патентов класса H01, выданных за 2012 г. У этих патентов в общей сложности насчитывается 650 цитируемых патентов прототипов, которые выданы начиная с 1994 г. Цитируемым патентом-прототипом для рассматриваемого патента является патент, который был выдан раньше рассматриваемого патента и который содержит положения, релевантные положениям в рассматриваемом патенте, в том числе положения, опровергающие часть новизны рассматриваемого патента. Поиск производился среди множества из всех патентов секции H и всех патентов из тех же подгрупп, что и цитируемые патенты. В общей сложности в рамках эксперимента учитывались описания порядка 50 тыс. патентов. Таким образом, подобранное множество относится к той же секции, что и тестовый набор заявок – это должно максимально затруднить поиск цитируемых патентов относящимся к другим секциям. Оценка качества разработанных методов производилась по методике, применяемой на соревнованиях в рамках CLEF [7]. В качестве показателей качества были взяты показатели "recall" для выборок из топ 1000, 500, 200, 300, 100, 50 наиболее релевантных найденных патентов. Показатель "recall" в данном случае показывает процент попадания цитируемых патентов прототипов в исходную выборку релевантных патентов. Показатель "recall" равный 100 % обозначает полное попадание всех цитируемых патентов прототипов в список наиболее релевантных найденных патентов. Результаты сравнения показателей "recall" разработанного метода, примененного для поиска всех цитируемых патентов-прототипов, разработанного метода, примененного для поиска только цитируемых патентов-прототипов, которые не относятся к секции H, и базового метода, основанного на мере TF-IDF, приведены в таблице.

Таблица

Результаты сравнения показателей «recall» для тестового набора текстов

Методы	Recall50	Recall100	Recall 200	Recall 300	Recall 500	Recall 1000
Разработанный метод	72	81	90	92	96	98
TF-IDF	32	44	50	53	58	66
Разработанный метод (цитаты, которые не принадлежат секции H)	55	60	76	80	85	90

Как можно видеть из таблицы, разработанный метод значительно превосходит традиционный. Даже поиск цитат из патентных классов, отличных от основного класса заявки, показывает высокие результаты. Однако такие цитаты составляют в среднем 10 % от общего количества цитат. Поэтому можно считать, что в большинстве случаев они не оказывают сильного влияния на общее качество поиска по уровню техники. Достигнутые показатели "recall" позволяют применять разработанный метод при проведении реальных патентных экспертиз.

В дальнейшем планируется применить методы машинного перевода для реализации поиска по уровню техники для поступившей заявки среди патентов на языках, отличных от языков заявок, поступивших в ФИПС.

Выражаем слова благодарности Всемирной организации интеллектуальной собственности и Роспатенту за информационную и финансовую поддержку.

Список литературы

1. D'hondt Eva. Combining Document Representations for Prior-art Retrieval / Eva D'hondt, Suzan Verberne, Wouter Alink, Roberto Cornacchia // CLEF Notebook Papers/Labs/Workshop. – 2011. – Режим доступа: http://sverberne.ruhosting.nl/papers/CLEF2011_workingnotes_dhondt_final.pdf, свободный. – Заглавие с экрана. – Яз. англ.

2. Graf E. Knowledge modeling in prior art search / E. Graf, I. Frommholz, M. Lalmas, K. Rijsbergen // Advances in Multidisciplinary Retrieval : First Information Retrieval Facility Conference. – Vienna : Springer, 2010. – P. 31–46.

3. Kravets Alla G. Enterprise Intellectual Capital Management by Social Learning Environment Implementation / Alla G. Kravets, Alexandr S. Gurtjakov, Anatoliy P. Darmanian // World Applied Sciences Journal. – 2013. – Vol. 23 (7). – P. 956–964.

4. Kravets Alla G. Corporate intellectual capital management: learning environment method / Alla G. Kravets, Alexandr Gurtjakov and Andrey Kravets // IADIS International Conference ICT, Society and Human Beings 2013 : part of the MCCSIS 2013 Conference. – Prague, 2013. – P. 3–10.

5. Magdy W. Applying the KISS Principle for the CLEF-IP 2010 Prior Art Candidate Patent Search Task / W. Magdy, G. J. F. Jones // Workshop of the Cross-Language Evaluation Forum, LABs and Workshops, Notebook Papers. – 2010. – Режим доступа: http://doras.dcu.ie/15834/1/Applying_ the_KISS_Principle_for_the_CLEF-_IP_2010.pdf, свободный. – Заглавие с экрана. – Яз. англ.

6. Mahdabi Parvaz. Report on the CLEF-IP 2011 Experiments: Exploring Patent Summarization / Parvaz Mahdabi, Linda Andersson, Allan Hanbury, and Fabio Crestani // CLEF Notebook Papers/Labs/Workshop. – Amsterdam, 2011. – Режим доступа: http://www.researchgate.net /publication/221159660_Report_on_the_CLEF-IP_2011_Experiments_Exploring_Patent_Summarization, свободный. – Заглавие с экрана. – Яз. англ.

7. Piroi Florina. CLEF-IP 2011: Retrieval in the Intellectual Property Domain / Florina Piroi, Mihai Lupu, Allan Hanbury, Veronika Zenz // CLEF 2011 Labs and Workshop, Notebook Papers. – Amsterdam, 2011. – Режим доступа: http://www.researchgate.net/publication/221159804_CLEF-IP_2011_Retrieval_in_the_Intellectual_ Property_Domain, свободный. – Заглавие с экрана. – Яз. англ.

8. Robertson S. E. Okapi at TREC-4 / S. E. Robertson, S. Walker, M. M. Beaulieu, M. Gatford, A. Payne // The 4th Text REtrieval Conference (TREC-4). – 1996. – Режим доступа: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.49.3311&rep=rep1&type=pdf, свободный. – Заглавие с экрана. – Яз. англ.

9. Segalovich Ilya. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine / Segalovich Ilya // MLMTA. – 2003. – Режим доступа: http://download.yandex.ru/company/iseg-las-vegas.pdf, свободный. – Заглавие с экрана. – Яз. англ.

10. Toutanova Kristina. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger / Kristina Toutanova, Christopher D. Manning // Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics. – Stroudsburg, 2000. – Vol. 13. – P. 63–70.

11. Verma Manisha. Exploring Keyphrase Extraction and IPC Classification Vectors for Prior Art Search / Manisha Verma, Vasudeva Varma // CLEF Notebook Papers/Labs/Workshop. – 2011. – Режим доступа: http://clef2011.clef-initiative.eu/resources/proceedings/Verma_Clef2011.pdf, свободный. – Заглавие с экрана. – Яз. англ.

12. WIPO Economics & Statistics Series. 2013 World Intellectual Property Indicators. – 2013. – Режим доступа: http://www.wipo.int/export/sites/www/freepublications/en/intproperty/941/wipo_pub_941_2013.pdf, свободный. – Заглавие с экрана. – Яз. англ.

13. Xue Xiaobing. Automatic query generation for patent search / Xiaobing Xue, W. Bruce Croft // The 18th ACM Conference on Information and knowledge management. – New York, 2009. – P. 2037–2040.

References

1. D'hondt Eva, Verberne Suzan, Alink Wouter, Cornacchia Roberto. Combining Document Representations for Prior-art Retrieval. *CLEF Notebook Papers/Labs/Workshop*, 2011. Available at: http://sverberne.ruhosting.nl/papers/CLEF2011_workingnotes_dhondt_final.pdf.

2. Graf E., Frommholz I., Lalmas M., Rijsbergen K. Knowledge modeling in prior art search. *Advances in Multidisciplinary Retrieval: First Information Retrieval Facility Conference*. Vienna, Springer, 2010, pp. 31–46.

3. Kravets Alla G., Gurtjakov Alexandr S., Darmanian Anatoliy P. Enterprise Intellectual Capital Management by Social Learning Environment Implementation. *World Applied Sciences Journal*, 2013, vol. 23 (7), pp. 956–964.

4. Kravets Alla G., Gurtjakov Alexandr, Kravets Andrey. Corporate intellectual capital management: learning environment method. *IADIS International Conference ICT, Society and Human Beings 2013: Part of the MCCSIS 2013 Conference)*. Prague, 2013, pp. 3–10.

5. Magdy W., Jones G. J. F. Applying the KISS Principle for the CLEF-IP 2010 Prior Art Candidate Patent Search Task. *Workshop of the Cross-Language Evaluation Forum, LABs and Workshops, Notebook Papers*, 2010. Available at: http://doras.dcu.ie/15834/1/Applying_the_KISS_Principle_for_the_CLEF-_IP_2010.pdf.

6. Mahdabi Parvaz, Andersson Linda, Hanbury Allan, Crestani Fabio. Report on the CLEF-IP 2011 Experiments: Exploring Patent Summarization. *CLEF Notebook Papers/Labs/Workshop*. Amsterdam, 2011. Available at: http://www.researchgate.net/publication/221159660_Report_on_the_CLEF-IP_2011_Experiments_Exploring_Patent_Summarization.

7. Piroi Florina, Lupu Mihai, Hanbury Allan, Zenz Veronika. CLEF-IP 2011: Retrieval in the Intellectual Property Domain. *CLEF 2011 Labs and Workshop, Notebook Papers*. Amsterdam, 2011. Available at: http://www.researchgate.net/publication/221159804_CLEF-IP_2011_Retrieval_in_the_Intellectual_Property_Domain.

8. Robertson S. E., Walker S., Beaulieu M. M., Gatford M.,. Payne A Okapi at TREC-4. *The 4th Text REtrieval Conference (TREC-4)*. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.49.3311&rep=rep1&type=pdf.

9. Segalovich Ilya. A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine. *MLMTA*, 2003. Available at: http://download.yandex.ru/company/iseg-las-vegas.pdf.

10. Toutanova Kristina, Manning Christopher D. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics*. Stroudsburg, 2000, vol. 13, pp. 63–70.

11. Verma Manisha, Varma Vasudeva. Exploring Keyphrase Extraction and IPC Classification Vectors for Prior Art Search. *CLEF Notebook Papers/Labs/Workshop*. 2011. Available at: http://clef2011.clef-initiative.eu/resources/proceedings/Verma_Clef2011.pdf.

12. *WIPO Economics & Statistics Series. 2013 World Intellectual Property Indicators.* 2013. Available at: http://www.wipo.int/export/sites/www/freepublications/en/intproperty /941/wipo_pub_941_2013.pdf.

13. Xue Xiaobing, Croft Bruce W. Automatic query generation for patent search. *The 18th ACM Conference on Information and Knowledge Management*. New York, 2009, pp. 2037–2040.