Evaluation of Retrieval Performance in Historical Newspaper Archives comparing Page-level and Article-level Granularity

Frank Buhr*, Bernd Neumann**

Abstract

The National Digital Newspaper Program (NDNP) has the goal to enhance the access to American historical newspapers for diverse user groups. One of the design decisions concerns the prestructuring which should be performed for archived newspapers. In this report we compare the retrieval performance for newspapers structured into pages ("page-level retrieval") with newspapers further decomposed into articles ("article-level retrieval"). The investigation is based on the complete1905 volume of The Washington Times comprising 2694 pages with 20800 articles. A set of 83 conjunctive-keyword queries has been used to determine the differences of precision and recall incurred by page-level retrieval as opposed to article-level retrieval. It is shown that page-level responses will on the average contain about 70% irrelevant hits due to keywords distributed over several semantically unrelated articles. Furthermore, hits may be missed if articles and keywords extend over more than one page. Examples illustrate that a more refined prestructuring of archived newspapers into components such as headlines, pictures, picture captions and advertisements will further improve retrieval performance. Disadvantages of page-level retrieval can be partly compensated by ranking procedures which help to concentrate relevant hits in the top part of a hitlist. However, this will only be an improvement for users who are not interested in all relevant hits of a hitlist.

1. Introduction

This report describes research carried out by Hamburger Informatik Technologie-Center (HITeC) for Content Conversion Specialists (CCS) in the context of the National Digital Newspaper Program (NDNP) funded by the National Endowment for the Humanities (NEH) and coordinated by the Library of Congress (LC). The main goal of the NDNP is to enhance the access to American historical newspapers for diverse user groups, including librarians, students, academic scholars and the general public. This goal will be achieved by providing web access to millions of digitized newspaper pages including directory information and search engines.

---

* buhr@hitec-hh.de, HITeC Hamburg, Germany
** neumann@informatik.uni-hamburg.de, Department for Informatics, Hamburg University, Germany

One of the design decisions of the NDNP is whether to prestructure digitized newspapers papers into articles and possibly further components such as headlines, pictures or advertisements. This would call for document analysis methods beyond standard OCR. CCS offers this advanced technology. Given an archive with newspapers structured at the article level, one can expect that query responses will be more precise than responses at page level. Consider, for example, a query where two keywords are entered in order to retrieve an article containing these two keywords. If structured at the page level, pages may be retrieved where the keywords belong to different articles and hence would not constitute useful responses. Furthermore, if articles extend over more than one page, they may not be retrieved at all at page level if the keywords are distributed on different pages.

To illustrate the possible drawback of page-level retrieval, consider a user who is interested in reports about steamship traffic on the Potomac in The Washington Times issues of the year 1905. She submits a Google-like query in terms of the two keywords "Potomac" and "steamship" and receives a hitlist of 161 pages. The first hit is shown in Fig. 1 which depicts a section of Page 2 of The Washington Times of August 28, 1905.

THE WASHINGTON TIMES, SUNDAY, AUGUST 20, 1905.

GRAND ARMY WILL CAMP IN DENVER

Plans for Annual Reunion Next Month.

WASHINGTON'S DELEGATION

Department of Potomac, in a Special, Will Travel to the West in Style.

The members of the Grand Army of the Republic, Department of the Potomac, have completed all arrangements for their journey to Denver, Col., to attend the thirty-ninth national encampment, which will be held in that city September 4, 5, 6, and 7.

A special train has been chartered for the trip from Washington to Denver.

The party which will represent the Department of the Potomac at the encampement will include many friends of the G. A. R. who are regular attendants at these encampments. Over one hundred have already applied for place on the train, which will be known as the "Department of the Potomac's Special," and it is expected that this number will be greatly augmented before September 1, when the party will leave for the encampment.

Will Have Four Delegates.

At Denver the Department of the Potomac will be represented by four delegates, who will each have one vote in the election. These delegates are Dan A. Grosvenor, Bernard T. Janney, John T. Layton, and H. M. Bennett. In addition to the four delegates the Depart-

VETERANS WHO WILL LEAD DELEGATION FROM WASHINGTON TO THE CONVENTION OF THE G. A. R. AT DENVER

L. VANDERHOEF, Assistant Adjutant General, Department of the Potomac.

JAMES TANNER, Candidate for the Office of Commander-in-Chief.

A. P. TASKER, Commander of the Department of the Potomac.

New Creation Theory By Chicago Savants

Eminent Professors of University Give to the World a Nebular Hypothesis Which No Doubt Will Attract Universal Attention.

CHICAGO, Aug. 19.—A new nebular hypothesis has been given to the world by two eminent professors of the University of Chicago. Frederick R. Moulton, which were to be found nuclei and local condensation. These gathered and formed the planets.

CONTRACT IMMIGRANTS SENT HERE IN DROVES

State Department Notified of Action Taken in Austria-Hungary by Steamship Companies.

The official correspondence between the State Department and the Austro-Hungarian foreign office, relating to governmental assistance being given undesirable immigrants under a contract with the steamship companies, was yes-

SUPREME COURT FREES TOO MANY?

Its Action Toward St. Elizabeth's Causes Comment.

DOCTORS CLASH OVER CASES

Local Physicians' Evidence Has More Weight With Court Than That of Hospital Doctors.

The frequency with which persons committed to St. Elizabeth's Hospital for the Insane because of their alleged insanity have recently been released from that institution through proceedings in the Supreme Court of the District is causing much comment. Within the past two months six persons, five soldiers and sailors and a civilian, have been released. Some of these have been discharged as the result of proceedings in habeas corpus instituted in their behalf and others as a result of a rehearing as to their mental condition.

In every instance some of the physicians attached to the Government Hospital—often one of them was Dr. W. A. White, superintendent of the institution—have given evidence tending to show that the subjects of the investigation were of unsound mind. In some instances their testimony was to the effect that the persons alleged to be insane were suffering from suicidal mania or delusional or persecutory insanity.

Evidence for Patients.

Against this evidence in favor of the person seeking release from the hospital

Fig. 1: Undesired response: page section with query keywords "Potomac" and "steamship" in different articles

The article in the left column contains several occurrences of "Potomac" (only one highlighted). It is about a journey – by train – of the "Grand Army of the Republic, Department of the Potomac". The article in the fourth column contains several occurrences of "steamship" (only one highlighted) and is about a contract between Austria-Hungary and Cunard Line to ship immigrants to the USA.

Obviously, due to the coarse page-level structure, an undesired response has been returned. None of the articles of this page contains both keywords, and the topics of the articles with one highlighted keyword are in fact unrelated to the query. Going down the hitlist, the user would find 157 more such non-hits. Fortunately, there are also 4 relevant hits, for example the one at Position 2 in the hitlist, shown in Fig. 2.

4 THE WASHINGTON TIMES, SUNDAY, AUGUST 6, 1905.

ALEXANDRIA FERRY BEGINS OCTOBER 1

First New Boat Will Be Launched at Wilmington, Del., in Near Future—New Day Boat for Old Point and Norfolk.

The new ferry between Washington and Alexandria will be in operation on or about October 1.

Authority for this announcement is John Callahan, general manager of the Norfolk and Washington Steamboat Company, into the control of which corporation the ferry line passed a few months ago. It was hoped that boats might run sooner, and with that end in view Mr. Callahan made efforts to buy a boat in New York, but was unable to do so.

The first new ferryboat will be launched at the Harlan & Hollingsworth yards in Wilmington, Del., about the middle of the present month and will be turned over to the company about the middle of September. The second boat to be placed on the line will be ready for service about five months later.

Fine Boats.

Both vessels will be constructed after the latest modern models and equipped in the best manner. The first boat will be named Woodbury, in honor of the president of the company, Levi Woodbury; for the other boat no name has as yet been selected.

The schedule to be installed calls for a round trip every hour, between 6 o'clock a. m. and 10 p. m., the time from city to city being twenty minutes. Efforts will be made to get a spare boat which would be put on the line until the second boat is completed. During the winter an arrangement will be made to enable Alexandrians to attend the Washington theaters, a boat to be run from here after the close of the performance.

The ferry house and slips here have been completed and are in shape for business whenever the boat comes from Wilmington. The building presents a bright appearance, the front being in pebble dash and white. The passenger entrance is in the center; that for teams at the east end, and the west entrance and slip are reserved for the use of the new day boat of the Norfolk and Washington line. The Alexandria terminal is about half finished.

Steamer Fireproof.

The new day boat, as Mr. Callahan declares, will be the finest steamer that ever plied on the Potomac river or Chesapeake bay. It will be fireproof from keel to hurricane deck and from stem to stern. It will cost $[illegible],000, and the contract calls for its delivery to the company so that the day trips to Old Point Comfort and Norfolk can be begun on May 1 next.

The minimum speed is to be eighteen miles an hour, the trip from Washington to Old Point to be made in nine hours. The boat is to be fitted with engines of the inclined compound type, and there are to be four Scotch boilers, all below deck. The capacity of the steamer will be 2,500 passengers.

In order to give every opportunity for promenading, for playroom for children, and such other amusements as are usually undertaken on an excursion steamer, both decks, fore and aft, will be cleared for playgrounds. There will be twenty drawing-rooms on the after deck, some of them furnished with beds for the accommodation of invalids or for sudden emergencies. Meals will be served either at table d'hote or a la carte.

Stimulates Building.

At what points along the Potomac landings will be made has not yet been determined, owing to the fact that certain parties intend to make extensive improvements at Lower Cedar Point and Piney Point, both favorite summer resorts for Washingtonians. It is also understood that, as a result of the installation of this day service on the river, a number of Washington people intend to build cottages at these points.

For years the desire has been expressed here in Washington that there might be a day service between this city and Old Point, because a day trip on the Potomac is a joy and an inspiration. There is little doubt, therefore, that the new venture of the steamboat company will enjoy ample patronage.

The officers of the Norfolk and Washington Steamboat Company are: Levi Woodbury, president; Clarence F. Norment, first vice president, and John Callahan, second vice president and general manager.

Mr. Callahan has been closely identified with the company for more than a quarter of a century and enjoys the reputation of being one of the most efficient managers of steamboat lines in the country. It was through his sagacity and foresight that the company obtained from the Legislature of Virginia and holds today three distinct charters by virtue of which it has the [illegible] to operate coastwise and ocean steamship lines, and build and operate [illegible] fifty miles inland from any point along its routes, a privilege which may prove of great advantage in the future.

KISS COST AGGRESSIVE HARVEY COPELAND $41

BUFFALO, Aug. 5.—Miss Eleanor J. Omphalius sued Harvey Copeland for $500 damages, alleging he had kissed and hugged her against her will.

The young woman is pretty, but Judge Hammond thought her valuation of one of her kisses excessive, and ordered Copeland to pay $85 damages and $9.20 costs.

BIASED.

"They say she made her husband what he is."

"H'm. Evidently she never had any previous experience," replied the old maid, who once thought she was going to get him.—Chicago Record-Herald.

WASHINGTON RIVER FRONT ADORNED BY NEW STRUCTURE

FERRY HOUSE AND SLIPS.
Just Completed by Norfolk and Washington Steamboat Company.

AUTOMOBILE'S SPEED TO BE PHOTOGRAPHED

Englishman's Ingenious Device Takes Pictures Which Will Make Evidence Against Chauffeurs Conclusive.

No more will the opulent member of the automobility and the humble but determined rural constable have a fibbing match over the speed at which the whiz car flew along the country road. No more will the chauffeur swear he was creeping at five knots an hour, while the deputy sheriff claims a mile a minute. These archaic situations are doomed to a place only in history.

Time photographs will in the future be flashed on the automobilist. With a sickly smile he will see absolute proof of his iniquity. The scheme has been evolved, according to a report to the Bureau of Manufactures, by an Englishman.

It consists of two time-recording synchronized cameras, set at the ends of a "trap." When the automobile breaks the string and snaps the first shutter, the camera photographs the automobile, its number, and its occupants, at the same time recording the instant of its passing.

The photograph taken by the second camera at the other end of the trap is similar, and by computing the difference in time, the authorities have the goods on the chauffeur. The recording device is so arranged that no tampering can be done, and the defendants must admit the accuracy of evidence.

"Look out for the cameras!" will now take the place of the old warning cry of the flying chauffeur. "Here are the constables!"

FEAR SCEPTIC POISONING FOR ISADOR WORMSER, JR.

SARATOGA, N. Y., Aug. 5.—Isador Wormser, jr., of New York, operated on for gallstones Thursday at the United States Hotel, does not seem to improve. While he has, according to Dr. C. S. May, recovered in a slight degree from the operation, there is danger of sceptic poisoning, which this forenoon threatens him.

CASE OF BLUFF.

"How did you succeed in accumulating such an immense fortune?" asked the lawyer who was drawing up the old man's will.

"By a very simple method," answered the aged millionaire. "When I was poor I pretended I was rich, and when I got rich I pretended I was poor."

ABSCONDER ALLEN ALIVE AND IN SOUTH AMERICA

Got Away With $100,000 of Preachers' Aid Society—Probably in Buenos Ayres.

BOSTON, Mass., Aug. 5.—That Willard S. Allen, of East Boston, treasurer of the Methodist Preachers' Aid Society, who absconded two years ago with more than $100,000 of the society's funds, is alive and well in South America, probably in Buenos Ayres, is the belief of the police here.

It was long believed that he had committed suicide. Chief Inspector Watts, who holds a warrant for Allen's arrest, said this morning that the police would redouble their efforts to effect Allen's arrest.

SHORE DINNER.

Bill—I guess that must have been a shore dinner we had yesterday, up at the house.

Jill—Why?

Bill—Well, there was sand in the sugar and salt water in the ice cream.—Yonkers Statesman.

Fig. 2: Desired response: page section with query keywords "Potomac" and "steamship" in a single article.

Here, the retrieved article is about launching a new ferry between Washington and Alexandria, related to the query of the user.

In this example, a newspaper archive structured at article-level would have saved the user the labour of inspecting and discarding 157 non-relevant pages, assuming that she was interested in articles containing all keywords.

In order to assess the gravity of the performance degradation by page-level retrieval, HITeC was asked to empirically compare the performance of newspaper information retrieval at article level and page level for a significant number of queries. The following sections present the approach taken for this investigation and the results obtained.

2. Approach to Performance Evaluation

2.1 Performance Measures

The main goal of the investigation was to provide objective measurements for usability and performance gains which can possibly be achieved by prestructuring digitized newspapers into components below the page level, such as articles, headlines, pictures, picture captions or advertisements. In view of the heterogeneous user groups addressed by the NDNP and the expected diversity of queries, we decided to restrict the performance evaluation to what we believe will be the most frequent type of queries: the search for articles based on keywords. To further simplify the task, only articles containing all keywords should be retrieved (different, for example, from internet search engines). As illustrated by the example in Section 1, prestructuring pages into articles prohibits that pages are retrieved where keywords are distributed over several articles. We assume now that such pages are irrelevant hits while pages with all keywords in at least one single article are relevant hits.

One way, then, to measure the performance difference between page-level retrieval and article-level retrieval is to compare the number of hits returned by either method for a given query. As all (almost, see Section 3.2) article-level hits will also be contained in the page-level hits, the difference of these numbers is the number of irrelevant hits as defined above.

A second performance criterion can be the position of a hit in a hitlist. Search engines usually return hitlists ranked by sophisticated algorithms. A ranking algorithm may perform at the page level quite different from the article level. Without knowledge of the ranking criteria, however, it is not possible to predict differences between page level and article level hitlist ordering.

2.2 Documents

For the evaluation it was necessary to have available a sufficiently large set of digitized newspapers, structured both at the page level and at the article level. It was decided that a complete volume of a particular newspaper would provide a sound basis, and we were provided with digitized versions of all issues of The Washington Times of the year 1905. The page-level and article-level XML annotations were generated by CCS conforming with the NDNP technical specifications.

The testbed files consisting of XML documents in ALTO and METS formats [2] were preprocessed and subsequently indexed by the open-source search engine Lucene [3, 4]. In order to compare page-level and article level data we created two types of records, *page records* and *article records*.

A page record consists of the plain text of one newspaper page and an URI that identifies the page. The plain text is generated by an XSLT-transformation from the METS physical structmap tag. The physical structmap tag references textblocks in the corresponding ALTO file (there is one ALTO file per page). All referenced textblocks are transformed to plain text and concatenated afterwards. The order of the textblocks is not changed.

An article record consists of the plain text of one newspaper article and a list of page URIs. The list of URIs denotes one or more pages on which the article appears. (One article might be continued on another page, so there is a one-to-many relationship between article records

and page records). The plain text of the article always starts with the heading of the article. Similarly to the processing of the page text, the article plain text is generated by an XSLT-transformation, but in this case the *logical* structmap of the METS file is used. The logical structmap represents a very fine grained view of a newspaper issue. Some of the main logical units are articles, headings, sections, subsections, paragraphs, illustrations, and captions. In order to generate each article's plain text, its substructures were flattened and all paragraphs containing text were concatenated (see Fig. 3a).

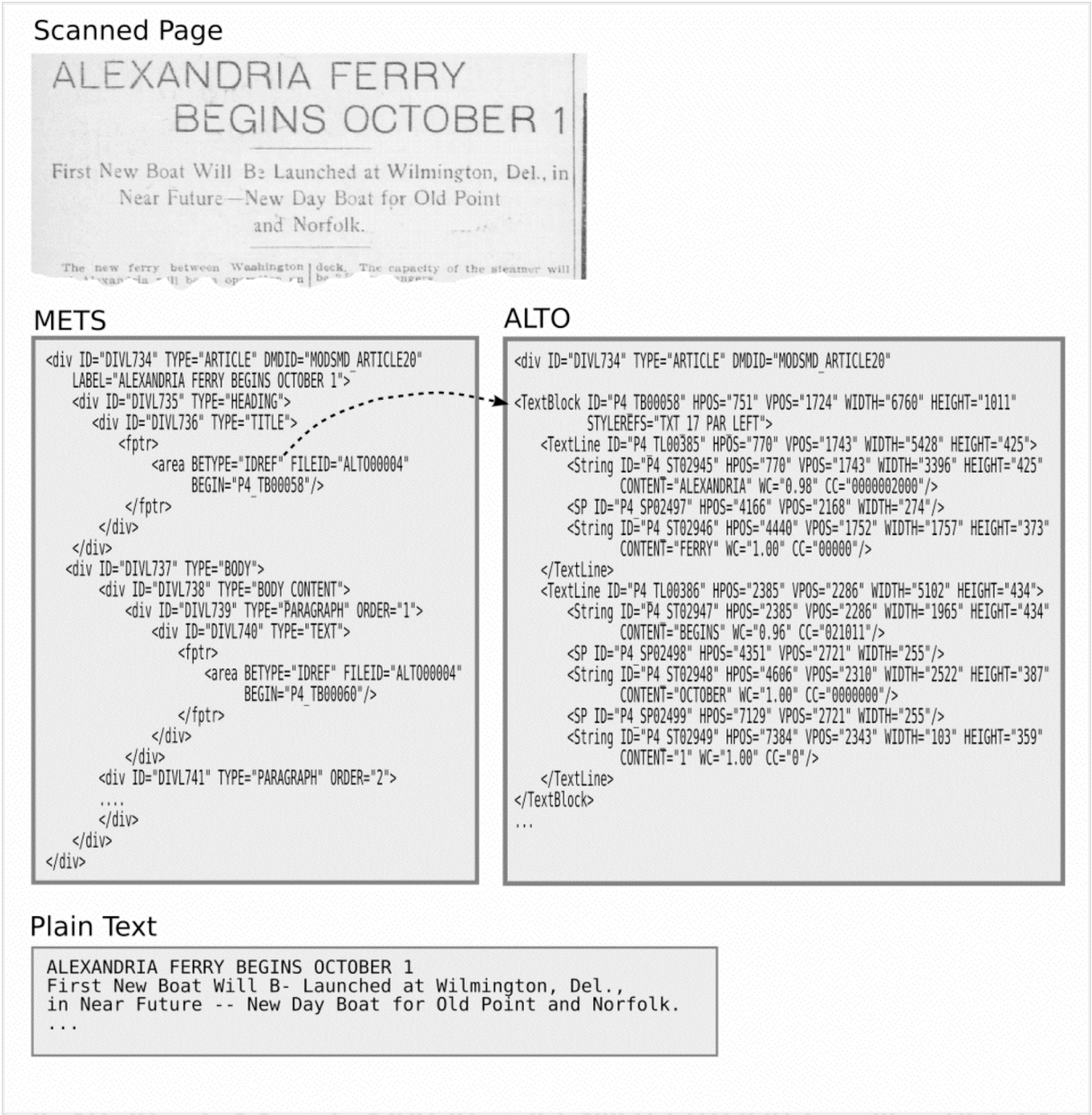


Fig. 3a: Plain text extraction. An article is represented logically in METS format. In order to extract the plain text, a link to a textblock inside the corresponding ALTO file is followed. The `CONTENT` of each `String` tag is concatenated. This is done for each logical unit (`BODY`, `BODY_CONTENT`, `PARAGRAPH`, etc.) in the METS structure.

The preprocessing resulted in a database of 2694 pages accessable for both page-level and article-level queries. The pages comprised 20800 articles with the article structure only accessible for article-level queries.

2.3 Queries

To familiarize ourselves with typical newspaper queries, we obtained concrete examples of retrieval requests submitted to librarians of the Library of Congress (Appendix A). It turned out that a large majority (33 out of 36) specified information at the directory level, e.g. specific dates of specific journals, which we could not satisfy with our experimental database consisting of the 1905 issues of The Washington Times. The strongly focussed queries may be due to the fact that search in all newspapers of the USA and extendable over an extensive time period (as envisioned by the NDNP) is currently inconceivable. Hence queries have been formulated most of which are quite limited in scope and not well suited for our evaluation.

We therefore prepared an own set of 83 queries, in the spirit of the examples, but expressed only with a small number of keywords in order to provoke responses with more than a single hit. The following are some examples of our list:

A "My grandfather has told me that the Potomac was frozen almost every year when he was young. Are there any reports confirming this ?"
Keywords: +Potomac +frozen

B "I wonder when firemen pension rights have been introduced. Are there newspaper reports about this?"
Keywords: +firemen pension +rights

C "My great-grandfather was lost on sea in a fisherboat during a storm on the atlantic in 1905. Is there a report about this?"
Keywords: +storm +atlantic

The complete list of keyword queries is given in the table below.

+potomac +frozen	+morgan +railway	+funeral +rothschild
+potomac +river +frozen	+morgan +McKinley	+funeral +miller
+potomac +steamship	+morgan +cab +accident	+storm +atlantic
+potomac +hampton	+taft +philippines	+henry +james
+japanese +immigrants	+panama +canal +immigrants	+indian +treaties
+open +door +china	+panama +canal +railroad	+ellis +island
+open +door +china +roosevelt	+big +stick +america	+morocco +roosevelt
+wilson +taft	+russia +japan	+entente +cordiale
+roosevelt +taft	+george +single +tax	+balfour +resign
+roosevelt +wilhelm	+firemen +pension	+protective +tariff
+hay +root	+firemen +pension +rights	+department +agriculture
+roosevelt +russia +japan	+smallpox +remedy	+department +justice
+refrigerator +fruit	+smallpox +cure	+morgan +h. +beach
+railway +fares +roosevelt	+smallpox +drug	+roosevelt +europe
+rockefeller +carnegie	+beef +trust	+new +york +church

+carnegie +gospel +wealth	+beef +trust +roosevelt	+southeast +washington
+gospel +wealth	+beef +trust +gorki	+memorial +services +miller
+carnegie +trade +union	+beef +trust +william	+memorial +services +jones
+carnegie +pacific +union	+beef +trust +wilhelm	+marriage +miller
+morgan +pacific +union	+suicide +student	+marriage +jones
+gulf +stream +activity	+fant +buried	+yacht +clubhouse
+secret +service +cotton	+jurors +coroner +murdered	+iowa +china
+prominent +clerk	+social +organization +louisville	+social +organisation +louisville
+afghanistan +troops +mission	+memorial +day +parade	+ambassador +mexico
+fire +life	+law +snow	+poor +boy +leading
+court +child	+norway +crisis +sweden	+roosevelt +texas +arrived
+suicide +student	+penalty +priest	+john +paul +jones
+north +pole +expedition	+atlantic +storm	

Table 1: Keyword queries used for performance evaluation. The "+" before each word marks a conjunctive query. (There is one duplicate query, "+suicide +student". This has no significant effect on the results)

2.4 Relevant Documents

The information about which documents in the collection are relevant in relation to a given query is crucial for the evaluation of retrieval performance. Only if the set of relevant documents for a query is known, the quality of retrieval results can be measured. In our experiment, we define the set of relevant page records for a query *q* in relation to the matching article records *Ha(q)* for the same query *q*. A matching page record for query *q* is called "relevant hit" if the page contains one or more articles from *Ha(q)*.

This is explained graphically in Fig. 3b. The query contains two keywords, W1 and W2 connected by a boolean AND operator (the '+'-Notation is just another syntax for that). In the left column, all matching article records *Ha(q)* are shown. The right column contains all matching page records. The edges between the article records and the page records depict the relationship "article appears on page". Page-level hit 1) contains one matching article, page-level hit 2) contains even two matching articles. Hence page-level hits 1) and 2) are both relevant hits. Page-level hit 3) does contain W1 and W2 but it does *not* contain a matching article. Therefore, page-level hit 3) is not a relevant hit.

What would have to be changed to make page hit 3) a relevant one? If page hit 3) had contained keywords W1 and W2 *together in one article*, this article would have been found by the article level query and would have appeared in the left column connected by an edge to the third page hit, hence in this case, page hit 3) it would have been considered a relevant hit.

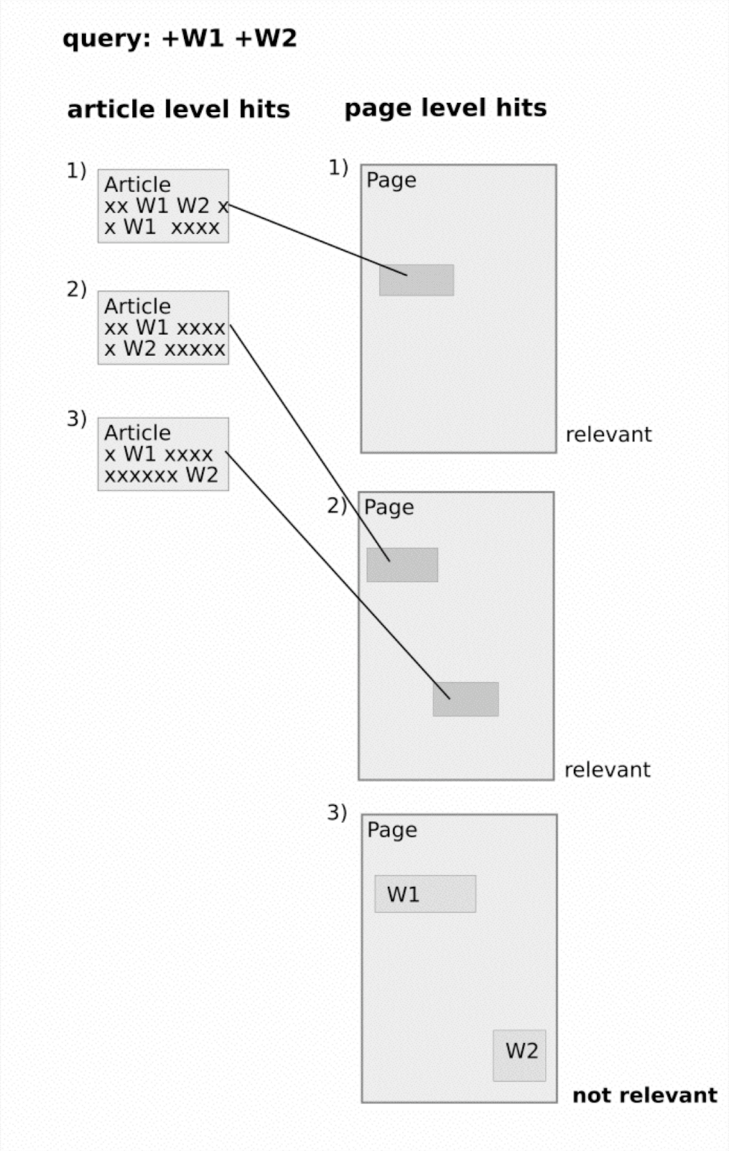


Fig. 3b: A page hit is considered a relevant hit if the page contains one or more hits of the article-level hitlist.

2.5 Search Engine

As already mentioned in Section 2.2, we used the open-source search engine Lucene [3, 4] for our retrieval experiments. The test framework has been programmed in Python. We used pyLucene [5] to get the "best of both worlds": the proven quality of Java-based Lucene and the flexibility of Python as an interactive, dynamically typed programming language.

We used basic Lucene classes (IndexWriter, SimpleAnalyzer, IndexSearcher, QueryParser) to build our retrieval framework. Ranking of search results is done by the standard score() Function, which implements a normalized tf-idf ranking (for boolean queries).

3. Results

3.1 Further Examples

We begin the performance comparison of page-level and article-level responses by inspecting further examples with queries taken from the list in Table 1. After studying the samples provided by Library of Congress (see Appendix A: Transcripts of real-life queries submitted to LC) we created a set of queries related to the year 1905. Each example we describe here begins with a fictional question an interested person might pose.

Example 1

I am a scientist at the University of Washington, Department of Atmospheric Sciences. For a study investigating the development of weather and especially hurricanes in the US states of the westcoast I am searching for interesting articles in past papers. Can You help me?

The query +gulf +stream +activity results in 7 page-level hits and 3 article-level hits.

The first hit at article level is very relevant as it deals with interesting information about gulf stream activitiy. All words appear in the headline and in the text as well (Fig. 4).

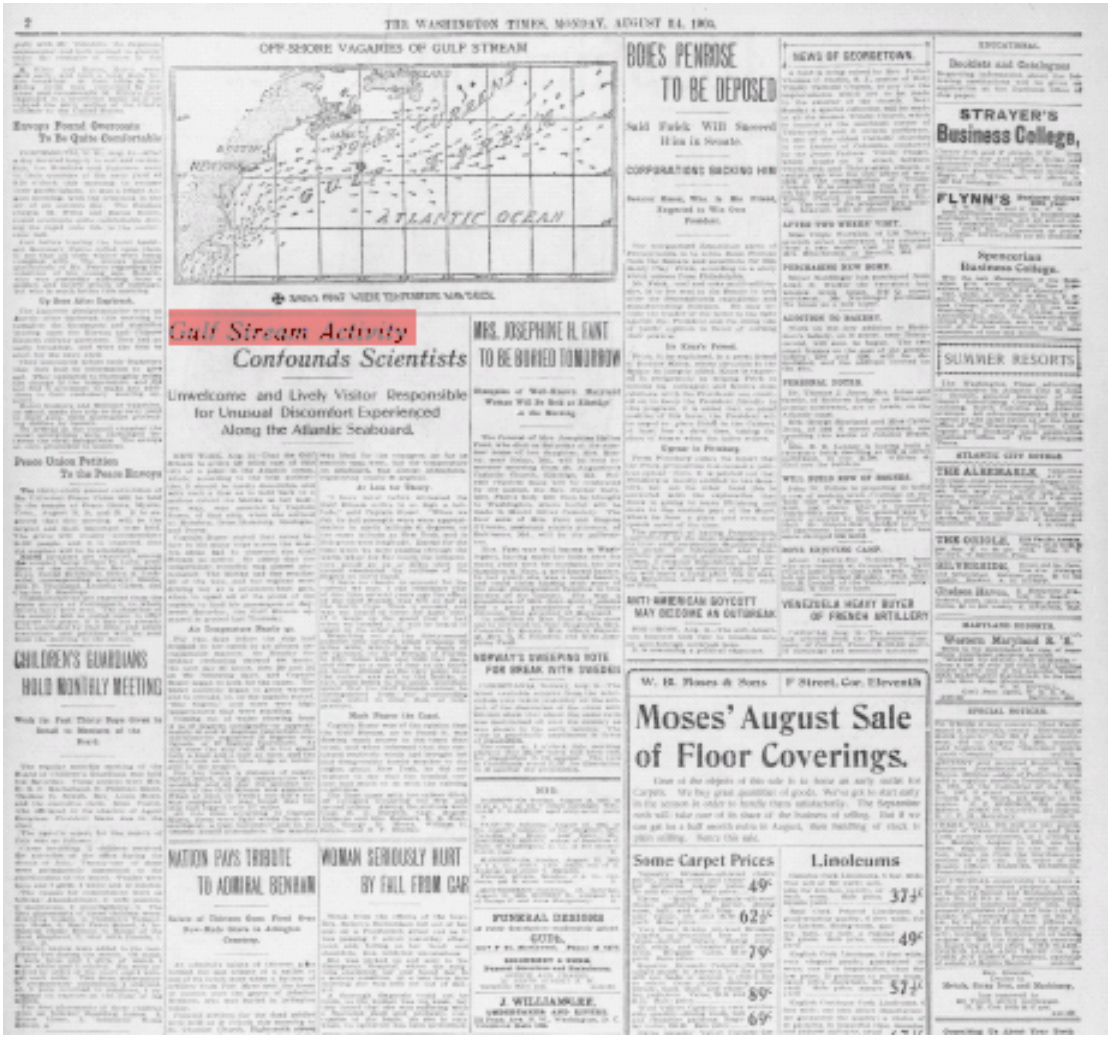
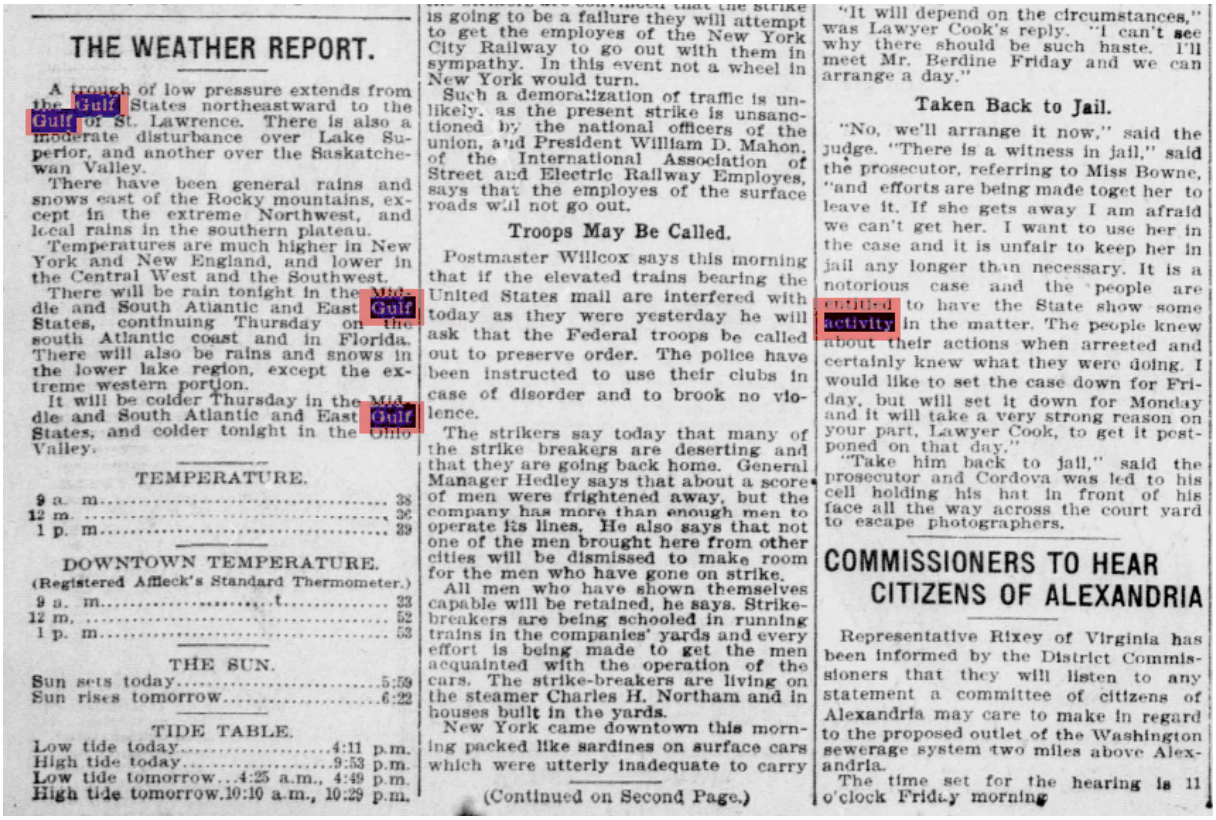


Fig. 4: First article-level hit for query +gulf +stream +activity.

Looking at the page-level hitlist we find the page in Fig. 5 at the third position of the hitlist. It shows a hit but each keyword appears in a different context.

Fig. 5: Third page-level hit for query +gulf +stream +activity.

The keyword “Gulf” appears a couple of times in the weather report, while the keyword “activity” is in a different article (see the close-up in Fig. 6).



THE WEATHER REPORT.

A trough of low pressure extends from the Gulf States northeastward to the Gulf of St. Lawrence. There is also a moderate disturbance over Lake Superior, and another over the Saskatchewan Valley.

There have been general rains and snows east of the Rocky mountains, except in the extreme Northwest, and local rains in the southern plateau.

Temperatures are much higher in New York and New England, and lower in the Central West and the Southwest.

There will be rain tonight in the Middle and South Atlantic and East Gulf States, continuing Thursday on the south Atlantic coast and in Florida. There will also be rains and snows in the lower lake region, except the extreme western portion.

It will be colder Thursday in the Middle and South Atlantic and East Gulf States, and colder tonight in the Ohio Valley.

TEMPERATURE.

9 a. m.	38
12 m.	36
1 p. m.	39

DOWNTOWN TEMPERATURE.
(Registered Affleck's Standard Thermometer.)

9 a. m.	33
12 m.	52
1 p. m.	53

THE SUN.

Sun sets today	5:59
Sun rises tomorrow	6:22

TIDE TABLE.

Low tide today	4:11 p.m.
High tide today	9:53 p.m.
Low tide tomorrow	4:25 a.m., 4:49 p.m.
High tide tomorrow	10:10 a.m., 10:29 p.m.

is going to be a failure they will attempt to get the employes of the New York City Railway to go out with them in sympathy. In this event not a wheel in New York would turn.

Such a demoralization of traffic is unlikely, as the present strike is unsanctioned by the national officers of the union, and President William D. Mahon, of the International Association of Street and Electric Railway Employes, says that the employes of the surface roads will not go out.

Troops May Be Called.

Postmaster Willcox says this morning that if the elevated trains bearing the United States mail are interfered with today as they were yesterday he will ask that the Federal troops be called out to preserve order. The police have been instructed to use their clubs in case of disorder and to brook no violence.

The strikers say today that many of the strike breakers are deserting and that they are going back home. General Manager Hedley says that about a score of men were frightened away, but the company has more than enough men to operate its lines. He also says that not one of the men brought here from other cities will be dismissed to make room for the men who have gone on strike.

All men who have shown themselves capable will be retained, he says. Strike-breakers are being schooled in running trains in the companies' yards and every effort is being made to get the men acquainted with the operation of the cars. The strike-breakers are living on the steamer Charles H. Northam and in houses built in the yards.

New York came downtown this morning packed like sardines on surface cars which were utterly inadequate to carry

(Continued on Second Page.)

“It will depend on the circumstances,” was Lawyer Cook's reply. “I can't see why there should be such haste. I'll meet Mr. Berdine Friday and we can arrange a day.”

Taken Back to Jail.

“No, we'll arrange it now,” said the judge. “There is a witness in jail,” said the prosecutor, referring to Miss Bowne, “and efforts are being made toget her to leave it. If she gets away I am afraid we can't get her. I want to use her in the case and it is unfair to keep her in jail any longer than necessary. It is a notorious case and the people are entitled to have the State show some activity in the matter. The people knew about their actions when arrested and certainly knew what they were doing. I would like to set the case down for Friday, but will set it down for Monday and it will take a very strong reason on your part, Lawyer Cook, to get it postponed on that day.”

“Take him back to jail,” said the prosecutor and Cordova was led to his cell holding his hat in front of his face all the way across the court yard to escape photographers.

COMMISSIONERS TO HEAR CITIZENS OF ALEXANDRIA

Representative Rixey of Virginia has been informed by the District Commissioners that they will listen to any statement a committee of citizens of Alexandria may care to make in regard to the proposed outlet of the Washington sewerage system two miles above Alexandria.

The time set for the hearing is 11 o'clock Friday morning

Fig. 6: Close-up of third page-level hit showing keywords in different articles.

Example 2

For a study in the class of political history at my college I am looking for a report in a US newspaper about the crisis between Norway and Sweden in the year 1905. Is it possible to get access to the LC research archive? I would like to enter for example a query with the keywords "norway" and "crisis" and "sweden".

The query results in 13 page-level hits and 8 article-level hits.

AROSE FROM LOWEST GRADE
IN THE REVENUE CUTTER SERVICE

CAPT. WORTH G. ROSS,
Recently Appointed by President Roosevelt as Chief of Important Bureau in Treasury Department as Reward of Faithful Service.

CAPTAIN ROSS MERITS HIS LATE PROMOTION

Long and Efficient Career of Officer Recently Appointed as Head of the Revenue Cutter Service.
Sketch of the Man.

Antecedent to the retirement of Capt. F. F. Shoemaker as chief of the Revenue Cutter Service, on April 1, President Roosevelt has appointed Capt. Worth G. Ross as his successor. Captain Ross is one of the most noted officers in the Revenue Cutter Service, and his promotion to the office which he now holds was in reward for faithful and con-

These men have been specially trained and educated for the service, and in the course of events must control its work and be held responsible for its efficiency for the future.

Captain Ross, however, represents no particular element in the service, as a large portion of the older officers are his warm personal friends and supporters, and it is confidently expected that his policy will be a broad and

NORWAY TO FIGHT WAR FOR LIBERTY

Alarm Sounded by Dr. Nansen, the Explorer.

WANTS HER INDEPENDENCE

Encroachments of Sweden, He Says, Will Yet Lead to Revolution.
Arctic Plans.

LONDON, April 17.—Dr. Nansen, the great Norwegian Arctic explorer, who is at present in London, seems to think that affairs between Sweden and Norway have reached a stage verging on revolt.

In the course of a letter, reviewing the historical side of the Scandinavian crisis, Dr. Nansen says:

"We are a peace-loving folk that, apart from our freedom, have no fonder desire than friendship with Sweden. But break into our house, try to hamper our freedom of action as a soverign state, and we—a self-respecting nation—have no choice in the matter, we must rise to the occasion and not willingly surrender our independence."

Arctic Exploration.

Continuing, he said that Arctic exploration was never further from his thoughts than at present, but that when the political crisis was over he might lead an expedition to the Antarctic.

"My whole mind is occupied with the useless crisis which has arisen between Norway and Sweden," he said. "I have no personal feeling, nor have Norwegians generally, against the Swedes, among whom I number many of my best friends, but we Norwegians feel that our independence is being encroached upon by Swedish politicians.

"We do not want to break the union of the two countries, but we wish to be assured that they are two independent kingdoms under one sovereign.

Fears for Future.

"Norway is determined to obtain a separate consular system, and should the royal sanction be refused I do not know what will happen.

"There are, of course, many other points of dispute, but that is one of the most important, and if it is not settled there is real danger of a disruption of the union.

"Norway has no intention of consulting or appealing to other powers on the question. It is not a matter for outside interference.

"I am not on a political mission in this country," added Dr. Nansen, laughingly.

"I am not an ambassador, but only the professor of a university."

Reverting to Polar research, Dr. Nansen expressed the opinion that the

Fig. 7: First article-level hit for query +norway +crisis +sweden.

The first hit at article level shows a very relevant article (Fig. 7). One of the keywords even appears in the headline.

The next article-level hit is relevant again (Fig. 8). Two of the keywords appear in the headline, the third keyword is further down in the article.

SWEDEN-NORWAY

TENSION ABATES

King Oscar Will Accept Address of Storthing.

MORE TALK OF REPUBLIC

Abandonment of the Monarchical Form of Government in Norway Meets With Approval.

(Special Copyright Cable.)

CHRISTIANIA, June 10.—The tension of the situation has been somewhat relieved by King Oscar's decision to accept the address of the storthing, which announced the dissolution of the union.

The refusal of the King to entertain the idea of permitting one of his sons to accept the Norwegian throne and the practical certainty that Prince Waldemar of Denmark will also be hindered from accepting, has revived the talk of a republic, which meets with general approval.

As one of the most prominent leaders in the secession movement said today: "While all our institutions at the present time are arranged in accordance with the monarchical form of government and a less drastic change would be required to arrange them for a constitutional monarchy than for a republic, the latter would be the natural government for the Norwegian people, who in olden times were thus governed, and who have maintained a strong spirit of independence through all the vicissitudes of the nation since it was merged with Denmark and Sweden under Eric in the fourteenth century."

FORMER BALTIMOREAN IS BURIED AT SEA

Priest

Procession

tigo, Wis

All Cla

Fig. 8: Second article-level hit for query +norway +crisis +sweden.

The page-level hitlist contains several hits which are not relevant at all. The hit shown in Fig. 9 contains the keywords “sweden” and “norway” at the bottom of the page while the keyword “crisis” is part of another article titled “Is Franco-Russion Allicance Ended?”.


NEWS AND CHAT FROM THE LEADING CAPITALS OF EUROPE

NEW ATTRACTION AT BISLEY CAMP

Two Lady Entries Add Interest to Event.

TENNIS AT WIMBLEDON

EXHIBIT FROM FAMOUS KRUPP GUN FACTORY AT ESSEN SHOWN AT BERLIN EXPOSITION

Is Franco-Russian Alliance Ended?

Compact Between Czar and German Kaiser Viewed Through Parisian Spectacles. Caustic Criticism of Allemanic Ruler.

Lion Fears Bear May Try to Embrace India

Soldiers Returning From Far East Could Be Diverted to Possible Conquest—Reform Probable in English Army.

DUCHESS WEDS A LION TRAINER

Paris Society Experiences a Unique Sensation.

FIERCE FIGHT WITH ESCAPED CONVICTS

Business Hours from 8 A. M. to 8 P. M., Saturdays Excepted

Summer Wrappers

59c

LANSBURGH & BRO

420-426 7th St. N.W.

Washable Dress Goods at One-third to One-half Less Than Usual Prices

A WIND-UP of the favored styles and patterns that have had a record of re-orders time and again to vouch for their popularity. They are as fresh as the lots that have preceded. But at a saving to you of a third to a half.

12½c for 36-inch Cannon Cloth

12½c for 19c Dotted Swiss

19c for 35c Organdie

25c for 37½c and 50c Organdies

12½c for 15c India Linon

8c for 12½c Lawns

25c and 35c Printed Dimities 19c

Galatea Suiting 16c

Clear-up of Summer Waists

Vacation Stationery

$2.98 $1.90 $2.90

Just Note the Prices on These Embroideries

Cool-Looking Bedspreads

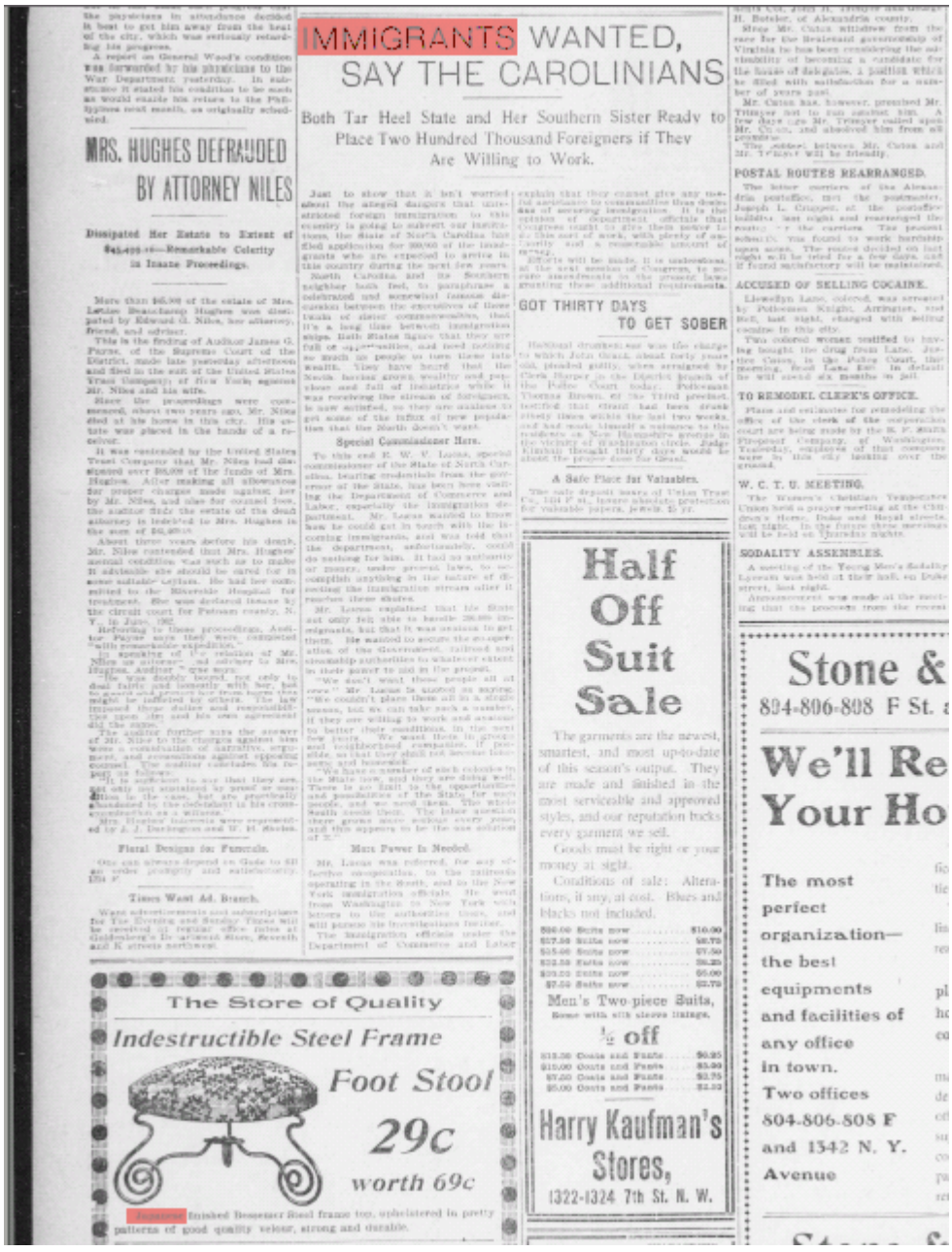
$1.25 $1.15 98c

Fig. 9: Irrelevant page-level hit for query +norway +crisis +sweden with keywords in different articles.

Example 3

My family came from Matsudo (Japan) to the USA when my grandfather was a baby. He often tells me about their problems in their early years. Now I am interested in the general situation of Japanese immigants to the US. I would like to search in the LC newspaper archive, is that possible? The public library in my small hometown in Virginia cannot provide such material.

The query +japanese +immigrants results in 15 page-level hits and 4 article-level hits.



IMMIGRANTS WANTED, SAY THE CAROLINIANS

Both Tar Heel State and Her Southern Sister Ready to Place Two Hundred Thousand Foreigners if They Are Willing to Work.

MRS. HUGHES DEFRAUDED BY ATTORNEY NILES

Dissipated Her Estate to Extent of $45,000.00—Remarkable Celerity in Insane Proceedings.

GOT THIRTY DAYS TO GET SOBER

Half Off Suit Sale

Harry Kaufman's Stores,
1322-1324 7th St. N. W.

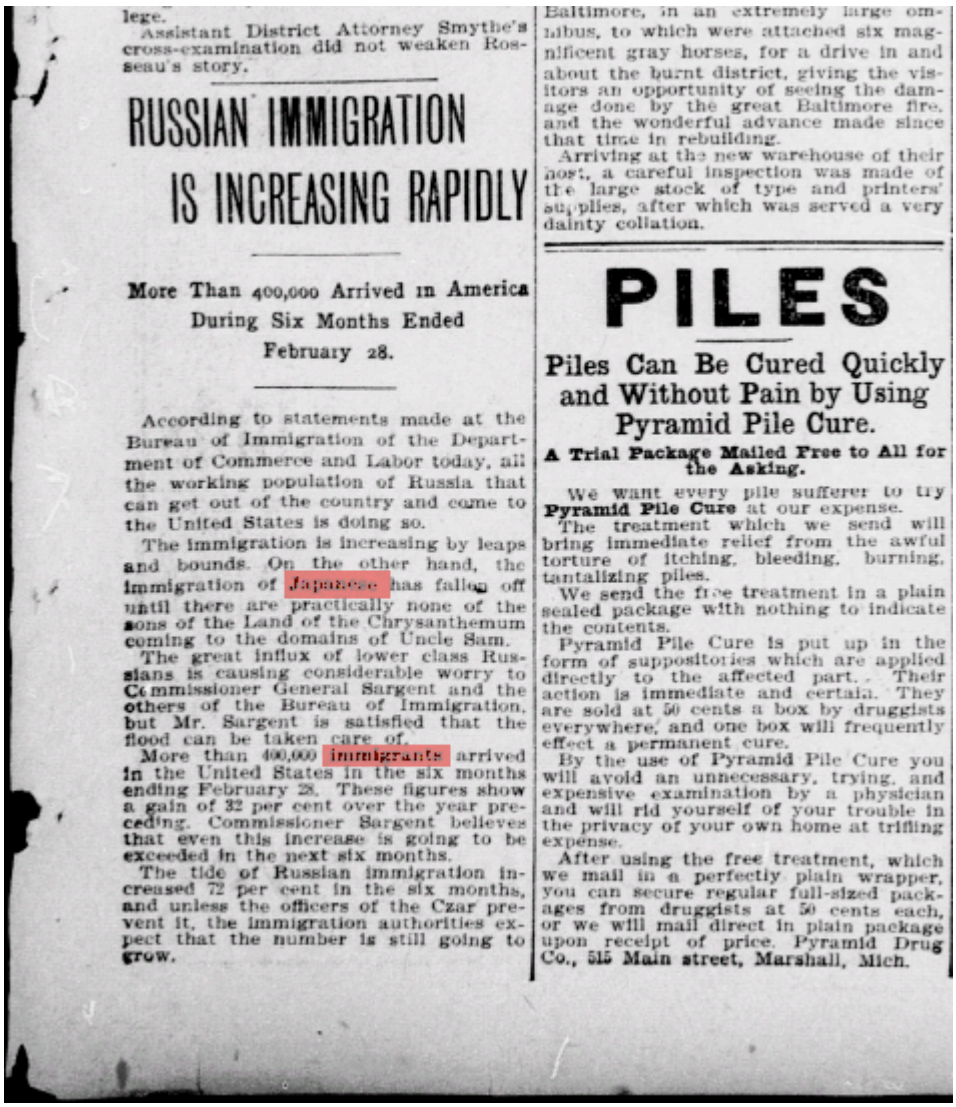
The Store of Quality
Indestructible Steel Frame
Foot Stool
29c
worth 69c

Stone &
804-806-808 F St.
We'll Re
Your Ho
The most perfect organization—the best equipments and facilities of any office in town.
Two offices 804-806-808 F and 1342 N. Y. Avenue

Fig. 10: Page-level hit for query +japanese +immigrants with one keyword in an article, the other in an advertisements.

Examining the page-level hitlist, we find the hit shown in Fig. 10. The keyword “immigrant” appears several times in the article ‘Immigrants Wanted Say the Carolinians’. The string “japanese”, however, appears in an advertisement at the bottom left of the page. This shows that it might be important to separate advertisements from article content.

A typical hit of the article-level hitlist is shown in Fig. 11.The article is clearly related to japanese (as well as russian) immigrants.



lege.

Assistant District Attorney Smythe's cross-examination did not weaken Rosseau's story.

RUSSIAN IMMIGRATION IS INCREASING RAPIDLY

More Than 400,000 Arrived in America During Six Months Ended February 28.

According to statements made at the Bureau of Immigration of the Department of Commerce and Labor today, all the working population of Russia that can get out of the country and come to the United States is doing so.

The immigration is increasing by leaps and bounds. On the other hand, the immigration of Japanese has fallen off until there are practically none of the sons of the Land of the Chrysanthemum coming to the domains of Uncle Sam.

The great influx of lower class Russians is causing considerable worry to Commissioner General Sargent and the others of the Bureau of Immigration, but Mr. Sargent is satisfied that the flood can be taken care of.

More than 400,000 immigrants arrived in the United States in the six months ending February 28. These figures show a gain of 32 per cent over the year preceding. Commissioner Sargent believes that even this increase is going to be exceeded in the next six months.

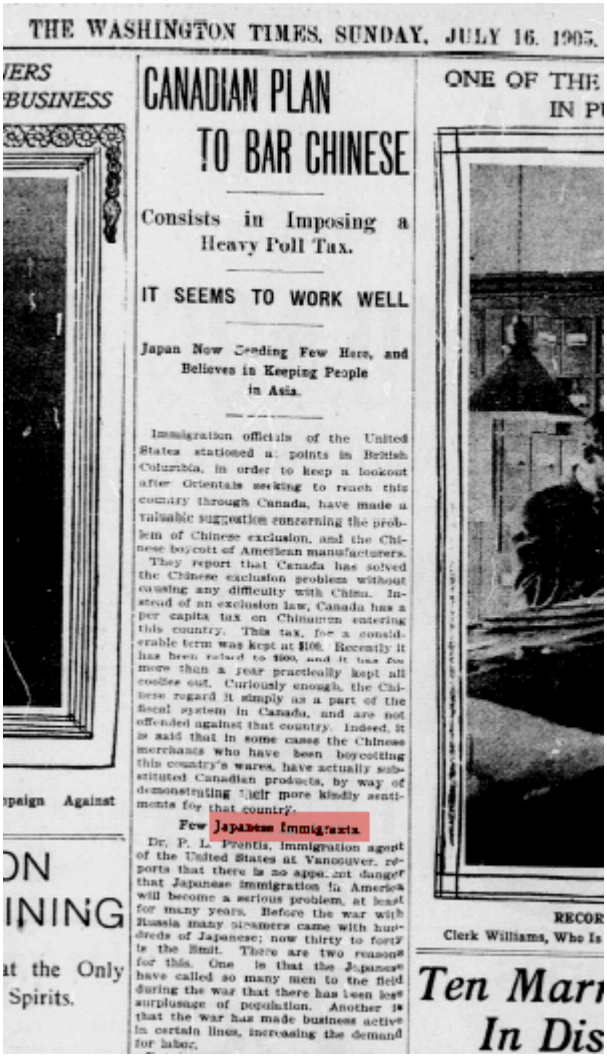
The tide of Russian immigration increased 72 per cent in the six months, and unless the officers of the Czar prevent it, the immigration authorities expect that the number is still going to grow.

Baltimore, in an extremely large omnibus, to which were attached six magnificent gray horses, for a drive in and about the burnt district, giving the visitors an opportunity of seeing the damage done by the great Baltimore fire, and the wonderful advance made since that time in rebuilding.

Arriving at the new warehouse of their host, a careful inspection was made of the large stock of type and printers' supplies, after which was served a very dainty collation.

Fig. 11: Article-level hit for query +japanese +immigrants with both keywords in a relevant article.

Another article-level hit with a meaningful result is shown in Fig. 12.



THE WASHINGTON TIMES, SUNDAY, JULY 16, 1905.

...ERS
...BUSINESS

...paign Against

...ON
...INING
...t the Only
Spirits.

CANADIAN PLAN TO BAR CHINESE

Consists in Imposing a Heavy Poll Tax.

IT SEEMS TO WORK WELL

Japan Now Sending Few Here, and Believes in Keeping People in Asia.

Immigration officials of the United States stationed at points in British Columbia, in order to keep a lookout after Orientals seeking to reach this country through Canada, have made a valuable suggestion concerning the problem of Chinese exclusion, and the Chinese boycott of American manufacturers. They report that Canada has solved the Chinese exclusion problem without causing any difficulty with China. Instead of an exclusion law, Canada has a per capita tax on Chinamen entering this country. This tax, for a considerable term was kept at $100. Recently it has been raised to $500, and it has for more than a year practically kept all coolies out. Curiously enough, the Chinese regard it simply as a part of the fiscal system in Canada, and are not offended against that country. Indeed, it is said that in some cases the Chinese merchants who have been boycotting this country's wares, have actually substituted Canadian products, by way of demonstrating their more kindly sentiments for that country.

Few Japanese Immigrants.

Dr. P. L. Prentis, immigration agent of the United States at Vancouver, reports that there is no apparent danger that Japanese immigration in America will become a serious problem, at least for many years. Before the war with Russia many steamers came with hundreds of Japanese; now thirty to forty is the limit. There are two reasons for this. One is that the Japanese have called so many men to the field during the war that there has been less surplusage of population. Another is that the war has made business active in certain lines, increasing the demand for labor.

ONE OF THE
IN P...

RECOR...
Clerk Williams, Who Is...

Ten Mar...
In Dis...

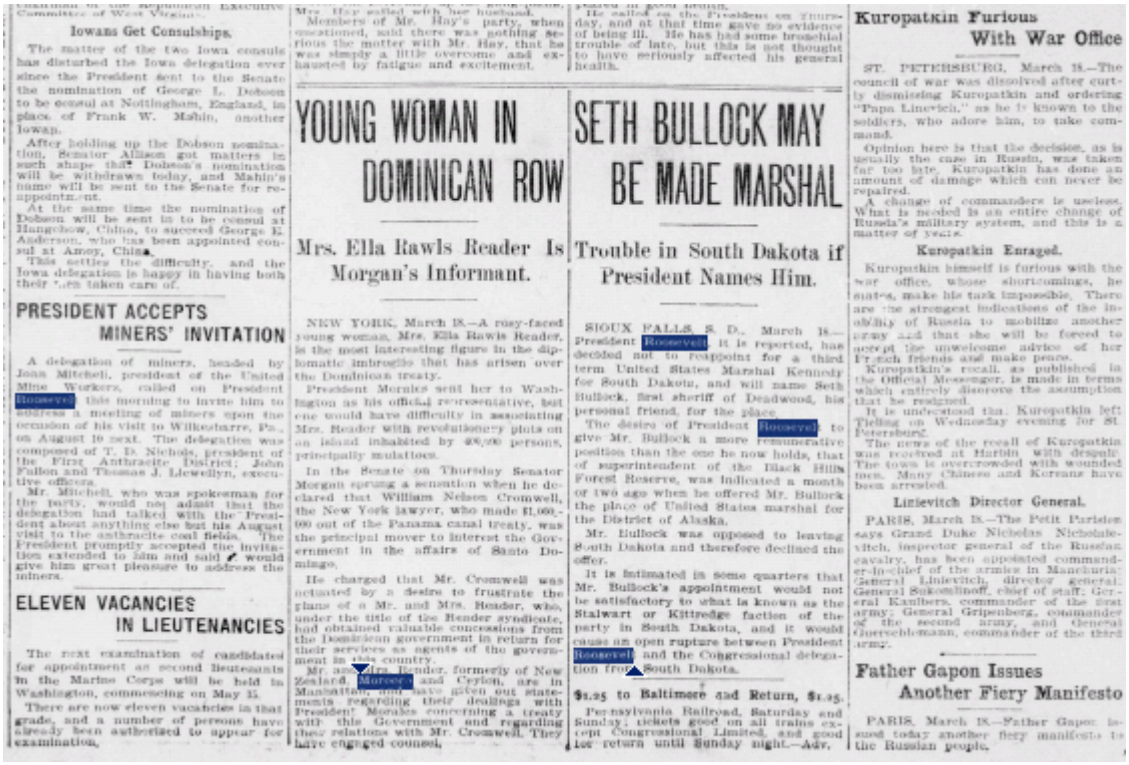
Fig. 12: Another article-level hit for the query +japanese +immigrants with both keywords in a relevant article.

Example 4

I am librarian in the McArthur Library in Biddeford (Maine). One user asked me to provide newspaper articles about the involvement and activities of President Roosevelt in the northern african state of Morocco.

The query +morocco +roosevelt results in 38 page-level hits and 5 article-level hits.

A page-level hit shows the keyword “roosevelt” several times in two different articles while “morocco” appears in a third article (Fig. 13). There is no relevant hit here as the two keywords do not appear in one meaningful context.



Iowans Get Consulships.

The matter of the two Iowa consuls has disturbed the Iowa delegation ever since the President sent to the Senate the nomination of George L. Dobson to be consul at Nottingham, England, in place of Frank W. Mahin, another Iowan.

After holding up the Dobson nomination, Senator Allison got matters in such shape that Dobson's nomination will be withdrawn today, and Mahin's name will be sent to the Senate for reappointment.

At the same time the nomination of Dobson will be sent in to be consul at Hangchow, China, to succeed George E. Anderson, who has been appointed consul at Amoy, China.

This settles the difficulty, and the Iowa delegation is happy in having both their men taken care of.

PRESIDENT ACCEPTS MINERS' INVITATION

A delegation of miners, headed by John Mitchell, president of the United Mine Workers, called on President Roosevelt this morning to invite him to address a meeting of miners upon the occasion of his visit to Wilkesbarre, Pa., on August 10 next. The delegation was composed of T. D. Nichols, president of the First Anthracite District; John Fallon and Thomas J. Llewellyn, executive officers.

Mr. Mitchell, who was spokesman for the party, would not admit that the delegation had talked with the President about anything else but his August visit to the anthracite coal fields. The President promptly accepted the invitation extended to him and said it would give him great pleasure to address the miners.

ELEVEN VACANCIES IN LIEUTENANCIES

The next examination of candidates for appointment as second lieutenants in the Marine Corps will be held in Washington, commencing on May 15.

There are now eleven vacancies in that grade, and a number of persons have already been authorized to appear for examination.

Mrs. Hay acted with her husband.

Members of Mr. Hay's party, when questioned, said there was nothing serious the matter with Mr. Hay, that he was simply a little overcome and exhausted by fatigue and excitement.

YOUNG WOMAN IN DOMINICAN ROW

Mrs. Ella Rawls Reader Is Morgan's Informant.

NEW YORK, March 18.—A rosy-faced young woman, Mrs. Ella Rawls Reader, is the most interesting figure in the diplomatic imbroglio that has arisen over the Dominican treaty.

President Morales sent her to Washington as his official representative, but one would have difficulty in associating Mrs. Reader with revolutionary plots on an island inhabited by 98,00 persons, principally mulattoes.

In the Senate on Thursday Senator Morgan sprung a sensation when he declared that William Nelson Cromwell, the New York lawyer, who made $1,000,000 out of the Panama canal treaty, was the principal mover to interest the Government in the affairs of Santo Domingo.

He charged that Mr. Cromwell was actuated by a desire to frustrate the plans of a Mr. and Mrs. Reader, who, under the title of the Reader syndicate, had obtained valuable concessions from the Dominican government in return for their services as agents of the government in this country.

Mr. and Mrs. Reader, formerly of New Zealand, Morocco and Ceylon, are in Manhattan, and have given out statements regarding their dealings with President Morales concerning a treaty with this Government and regarding their relations with Mr. Cromwell. They have engaged counsel.

SETH BULLOCK MAY BE MADE MARSHAL

Trouble in South Dakota if President Names Him.

SIOUX FALLS, S. D., March 18.—President Roosevelt, it is reported, has decided not to reappoint for a third term United States Marshal Kennedy for South Dakota, and will name Seth Bullock, first sheriff of Deadwood, his personal friend, for the place.

The desire of President Roosevelt to give Mr. Bullock a more remunerative position than the one he now holds, that of superintendent of the Black Hills Forest Reserve, was indicated a month or two ago when he offered Mr. Bullock the place of United States marshal for the District of Alaska.

Mr. Bullock was opposed to leaving South Dakota and therefore declined the offer.

It is intimated in some quarters that Mr. Bullock's appointment would not be satisfactory to what is known as the Stalwart or Kittredge faction of the party in South Dakota, and it would cause an open rupture between President Roosevelt and the Congressional delegation from South Dakota.

Kuropatkin Furious With War Office

ST. PETERSBURG, March 18.—The council of war was dissolved after curtly dismissing Kuropatkin and ordering "Papa Linevich," as he is known to the soldiers, who adore him, to take command.

Opinion here is that the decision, as is usually the case in Russia, was taken far too late. Kuropatkin has done an amount of damage which can never be repaired.

A change of commanders is useless. What is needed is an entire change of Russia's military system, and this is a matter of years.

Kuropatkin Enraged.

Kuropatkin himself is furious with the war office, whose shortcomings, he states, make his task impossible. There are the strongest indications of the inability of Russia to mobilize another army and that she will be forced to accept the unwelcome advice of her friends and make peace.

Kuropatkin's recall, as published in the Official Messenger, is made in terms which entirely disprove the assumption that he resigned.

It is understood that Kuropatkin left Tieling on Wednesday evening for St. Petersburg.

The news of the recall of Kuropatkin was received at Harbin with despair. The town is overcrowded with wounded men. Many Chinese and Koreans have been arrested.

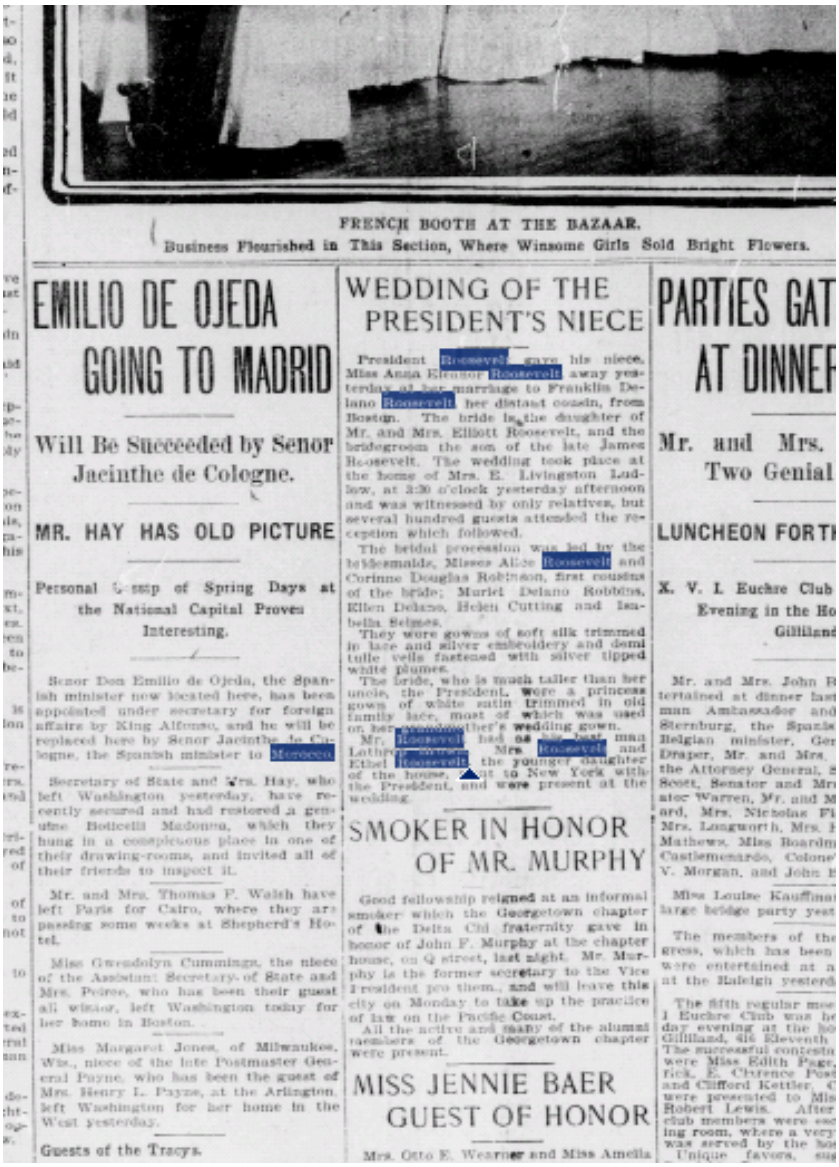
Linievitch Director General.

PARIS, March 18.—The Petit Parisien says Grand Duke Nicholas Nicholaievitch, inspector general of the Russian cavalry, has been appointed commander-in-chief of the armies in Manchuria; General Linievitch, director general; General Sukomlinoff, chief of staff; General Kaulbers, commander of the first army; General Gripenberg, commander of the second army, and General Guerschelmann, commander of the third army.

Father Gapon Issues Another Fiery Manifesto

PARIS, March 18.—Father Gapon issued today another fiery manifesto to the Russian people.

Another page-level hit (Fig. 14) shows “roosevelt” many times in an article about the wedding of the President’s niece while “morocco” appears in an article dealing with a very different topic.



FRENCH BOOTH AT THE BAZAAR.
Business Flourished in This Section, Where Winsome Girls Sold Bright Flowers.

EMILIO DE OJEDA GOING TO MADRID

Will Be Succeeded by Senor Jacinthe de Cologne.

MR. HAY HAS OLD PICTURE

Personal Gossip of Spring Days at the National Capital Proves Interesting.

Senor Don Emilio de Ojeda, the Spanish minister now located here, has been appointed under secretary for foreign affairs by King Alfonso, and he will be replaced here by Senor Jacinthe de Cologne, the Spanish minister to Morocco.

Secretary of State and Mrs. Hay, who left Washington yesterday, have recently secured and had restored a genuine Botticelli Madonna, which they hung in a conspicuous place in one of their drawing-rooms, and invited all of their friends to inspect it.

Mr. and Mrs. Thomas F. Walsh have left Paris for Cairo, where they are passing some weeks at Shepherd's Hotel.

Miss Gwendolyn Cummings, the niece of the Assistant Secretary of State and Mrs. Peirce, who has been their guest all winter, left Washington today for her home in Boston.

Miss Margaret Jones, of Milwaukee, Wis., niece of the late Postmaster General Payne, who has been the guest of Mrs. Henry L. Payne, at the Arlington, left Washington for her home in the West yesterday.

Guests of the Tracys.

WEDDING OF THE PRESIDENT'S NIECE

President Roosevelt gave his niece, Miss Anna Eleanor Roosevelt away yesterday at her marriage to Franklin Delano Roosevelt, her distant cousin, from Boston. The bride is the daughter of Mr. and Mrs. Elliott Roosevelt, and the bridegroom the son of the late James Roosevelt. The wedding took place at the home of Mrs. E. Livingston Ludlow, at 3:30 o'clock yesterday afternoon and was witnessed by only relatives, but several hundred guests attended the reception which followed.

The bridal procession was led by the bridesmaids, Misses Alice Roosevelt and Corinne Douglas Robinson, first cousins of the bride; Muriel Delano Robbins, Ellen Delano, Helen Cutting and Isabella Selmes.

They were gowns of soft silk trimmed in lace and silver embroidery and demi tulle veils fastened with silver tipped white plumes.

The bride, who is much taller than her uncle, the President, wore a princess gown of white satin trimmed in old family lace, most of which was used on her mother's wedding gown.

Mr. Roosevelt had as his best man Lothrop Motley. Mrs. Roosevelt and Ethel Roosevelt, the younger daughter of the house, went to New York with the President, and were present at the wedding.

SMOKER IN HONOR OF MR. MURPHY

Good fellowship reigned at an informal smoker which the Georgetown chapter of the Delta Chi fraternity gave in honor of John F. Murphy at the chapter house, on Q street, last night. Mr. Murphy is the former secretary to the Vice President pro tem., and will leave this city on Monday to take up the practice of law on the Pacific Coast.

All the active and many of the alumni members of the Georgetown chapter were present.

MISS JENNIE BAER GUEST OF HONOR

Mrs. Otto E. Wearner and Miss Amelia

PARTIES GAT AT DINNER

Mr. and Mrs. Two Genial

LUNCHEON FORT

X. Y. I. Euchre Club Evening in the Ho Gilliland

Mr. and Mrs. John B tertained at dinner last man Ambassador and Sternburg, the Spanis Belgian minister, Ger Draper, Mr. and Mrs. the Attorney General, S Scott, Senator and Mr ator Warren, Mr. and M ard, Mrs. Nicholas Fl Mrs. Longworth, Mrs. H Mathews, Miss Boardm Castlemenardo, Colonel V. Morgan, and John R

Miss Louise Kauffma large bridge party yest

The members of the gress, which has been were entertained at a at the Raleigh yesterd

The fifth regular mee I Euchre Club was he day evening at the ho Gilliland, 416 Eleventh The successful contestan were Miss Edith Page, rick, E. Clarence Post and Clifford Kettler. C were presented to Mis Robert Lewis. After club members were es ing room, where a very was served by the ho Unique favors, sui

Fig. 14: Another irrelevant page-level hit for query +morocco +roosevelt.

These examples show that irrelevant hits at page-level appear to be quite common. In comparison, article-level hits with all keywords in one article tend to meet the intentions behind keyword-based queries.

3.2 Summary of Retrieval Results for Queries

All queries listed in Table1 have been evaluated with all 2694 pages of the year 1905 volume of The Washington Times, first at page level, second at article level. Results are shown in the

table of Appendix B in detail. In this section we present summary information and provide some explanations.

Total number of page-level hits (sum of p over all queries)	9980
Total number of article-level hits (sum of a over all queries)	3210
Total number of pages with relevant hits (sum of pa over all queries)	3017
Mean fraction of relevant-hit pages in page-level hits (mean taken over the resulting pa/p of all queries)	0.28

We remind that all article-level hits are considered relevant hits (cf. Section 2.4). The results show that the large number of irrelevant hits in the page-level response of our examples is more the rule than the exception. On the average we may expect that between 65% and 75% of the page-level hits do not contain relevant information.

One may wonder whether this small fraction of relevant hits is true for virtually all page-level responses or whether this phenomenon varies, i.e. whether some responses have a large fraction of relevant hits and some have a very small one, even considerably smaller than the average of 30%. This can be shown by determining the fraction of relevant hits for each query response and counting fractions occurring in the intervals 0 to 0.1, 0.1 to 0.2, etc. The result is illustrated in the histogram presented in Fig. 15. It shows that in the large majority of cases, the fractions of relevant hits are between 0 and 0.4. Hence satisfying results with large fractions of relevant hits are very seldom.

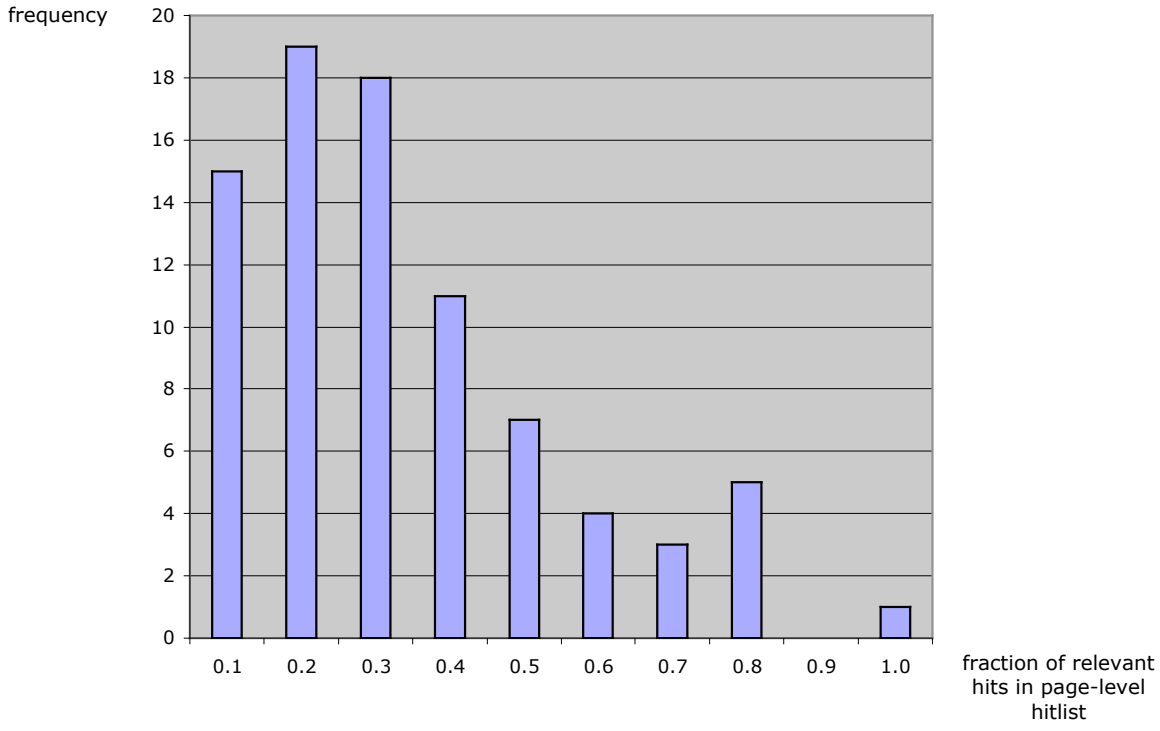


Fig. 15: Histogram of fractions of relevant hits in page-level hitlists.

The table in Appendix B also shows that for several queries a single page contains more than one relevant article (a > pa). Interestingly, there are also pages where pa is larger than a (e.g. for Query 65). In such cases not all article-level hits are contained in the corresponding page-level hitlists. This is due to the fact that articles may extend over more than one page. The CCS software used for prestructuring pages into articles effectively merges all article parts. Thus, keywords distributed over several pages are brought together and may allow hits at article level which would fail at page level.

We have determined the number of multiple-page articles for the1905 volume of The Washington Times:

Total number of articles	20800
Articles extending over 1 page	20369
Articles extending over 2 pages	428
Articles extending over 3 pages	2
Articles extending over 4 pages	1

Hence multiple-page articles constitute a fraction of 2%. This number may be considered small, but it is important to realize that multiple-page articles are completely missed by page-level retrieval. A missed hit is certainly much more undesirable than an irrelevant hit as the latter can be removed by inspection.

3.3 Retrieval Performance using Proximity Search

Our evaluation of retrieval performance in Section 3.2 is based on the effect of article structure on document granularity. Assuming that users are mainly interested in the semantic unit "article", we showed that retrieval performance degrades if the basic units of retrieval are large pages rather than smaller articles.

In this section we are going to answer the question: "Can page-level retrieval performance be enhanced without using article-level information?". The idea is that if fine-grained structures like articles are not explicitly represented in the document collection, it might nevertheless be possible to use a dynamic retrieval-time technique to lower the granularity.

Lucene provides a retrieval operator named *phrase query* that can be used for that.

A phrase query implements a proximity search. Given two keywords W1 and W2 and a distance *d*, we find documents containing words W1 and W2 in at most a distance of d words. This technique is commonly used to implement a fuzzy search predicate for composite names that can vary in the order of occurance of their parts (e.g. "Franklin Roosevelt", "Roosevelt, Franklin", "Franklin D. Roosevelt" and so on). We use much larger distances (several hundred words) to get an effect similar to a sliding window across a page.

Lucene's phrase query implementation has another important property: The distance of the occurences of the search terms influences the ranking of results. The closer two matching search terms occur on a page, the higher is their position in the hitlist. Thus phrase queries can be used to improve the ranking.

Fig. 16 shows precision-recall graphs (averaged over all 83 queries) for three conditions: without phrase queries (c0), with phrase queries and distance 50 (c50), and with phrase queries and distance 100000 (c100k).

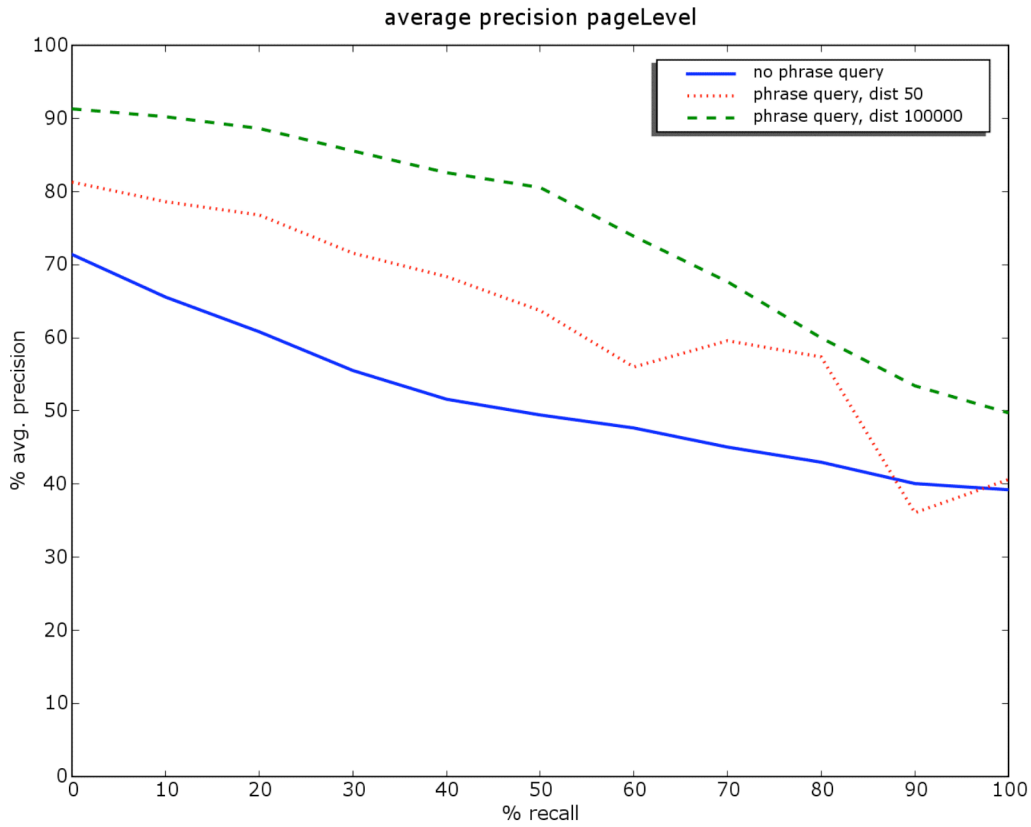


Fig. 16: Precision-recall graph. It shows the average precision over all queries in relation to percentage recall for standard boolean search ("no phrase queries") and phrase queries with distance 50 and 100000.

A precision-recall graph for a single query is constructed as follows (cf. Chapter 3 in [1]). Going down the list of ranked search results, at each relevant hit (at position n), the fraction of the number of relevant hits so far and the over-all number of hits so far, i.e. the precision-at-n, is computed. The values for n are normalized by the number of relevant documents (hence we get "percentage recall" for the x-axis). Further, the graph is interpolated, so that the x-values lie at 10% steps (0%, 10%, 20% and so on). A precision-recall graph for all queries is constructed by computing the average precision at each recall level.

The precision-recall graph at Fig. 16 shows a higher precision across the first recall levels for c50 and across all recall levels for c100k compared with c0.

What can we conclude from our phrase-query experiment? Clearly, the average precision can be significantly increased by using phrase queries. We get the highest increase with a very large distance of 100000. This is beyond the length (in number of words) of all articles and even of all pages. Hence, the improvement of performance must be due to the fact that phrase

queries provide a much better ranking than the standard Lucene ranking algorithm, which is tf*idf[*].

If users are expected to read only the first few hits, this increase in precision at the top of the list of search results is a clear benefit. If users are interested in the whole list of search results, the benefit is limited. In this case the precision recall graph in precRec1 does not tell the whole story, because "precision at 100% recall" means "precision at the last relevant hit". So Fig. 16 does not show the number of non-relevant hits below the last relevant hit – but a user interested in all hits would have to read them as well. For condition c100k we get as many hits as there are page hits in our first experiment in Section 3.2. Hence for users interested in a complete list of search results our results in Fig. 15 apply.

An important caveat for our phrase query experiment must be mentioned: We did not control the order of text blocks in the page-level plain text. We took the text blocks as they appeared in the ALTO files of our text collection. This is what would probably be done in a simple implementation of a fulltext search engine for page-level data. A pre-ordering, done for example by OCR processing, might lead to an ordering biased towards article-like clusters. On the other hand, OCR processing might as well cause text blocks to be torn apart which would be closer together if a simple column-wise ordering had been applied.

Another important aspect regarding the interpretation of Fig. 16 is the effect of only analyzing average query results. Standard deviation of precision (across queries) at 0% recall is 32% for condition c0 and 21% for condition c100k, but the average precision values at 0% recall differ only by 20%. Hence, ranking by phrase queries most likely cannot guarantee a better precision for all queries.

4. Conclusions

In this study we investigated the performance differences between historical newspaper information retrieval at page level and at article level. Our experiments were restricted to conjunctive-keyword queries which are in common use in search engines such as Google. Our results are based on a complete volume of The Washington Times of the year 1905.

We showed by a statistical analysis based on 83 queries that on the average 70% of a page-level hitlist is irrelevant in the sense that in such hits the keywords are distributed over several semantically unrelated articles. Hence, they do not comply with the presumed intention of a user to only retrieve articles containing all keywords.

For users who have to examine a large number of responses, this lack of precision will certainly be felt as a severe performance deficiency.

Page-level retrieval also gives rise to limited recall (i.e. to missed hits) if one article and the keywords are distributed over several pages. Such articles will not be returned as hits by page-level retrieval. If, however, the newspapers are structured into semantically coherent articles irrespective of page breaks (as has been done by the CCS software used in our experiments), this performance deficiency can be avoided. For The Washington Times of 1905, the fraction of multiple-page articles is 2%. This fraction may vary, of course, as other newspapers may have other layout styles.

[*] tf*idf (term frequency times inverse document frequency) is a standard ranking algorithm that yields higher scores for documents the more search terms occur in them and the less often documents with those terms occur in the whole collection.

We also investigated the distribution of relevant hits in page-level hitlists. If the relevant hits tend to be at the beginning of a page-level hitlist and spurious hits more towards the end, then the inconvenience of the lack of precision might be tolerable for many users. The distribution depends on the ranking procedure, of course, which may differ considerably between search engines. Our results show that for our data the default ranking in Lucene indeed tends to place relevant hits more towards the top of a page-level hitlist, thus yielding an average precision for the top entries of 70%. Another ranking procedure, phrase-query ranking, increases the average precision in the top of the hitlist even further to almost 90%. Of course, this is only an advantage for users who are happy with the first few hits and do not have to examine all relevant hits. We also showed that for a considerable fraction of queries a close-to-average performance cannot be expected due to the large variance of this ranking effect.

From our experiments we therefore conclude that prestructuring newspapers into articles will significantly improve the retrieval performance as compared to page-level retrieval.

The use cases provided by LC and the examples examined in The Washington Times issues of 1905 also illustrate that a further decomposition into headlines, pictures, picture captions, advertisements and other units may be very useful. As illustrated in Fig. 10, a page-level hit may even connect articles with advertisements if keywords are recognized accordingly. This - usually undesirable - effect can be avoided by refined prestructuring. Furthermore, if a user can specify that her query, or certain keywords, should be restricted to headlines or picture captions, for example, search may be narrowed down quite effectively. After all, the historical newspaper archive will contain millions of pages, and high retrieval performance will be mandatory if this service is to be accepted by the envisioned user groups.

References

[1] Baeza-Yates, R., Ribeiro-Neto, B. *Modern Information Retrieval.* ACM Press / Addison-Wesley, 1999.

[2] CCS GmbH. *METS / ALTO XML Object Model.* http://www.ccs-gmbh.com/alto/general.html

[3] Apache Lucene Project. *Lucene Search Engine.* http://lucene.apache.org/

[4] Gospodnetic, O., Hatcher, E. *Lucene in Action.* Manning Publications, 2005.

[5] *PyLucene.* http://pylucene.osafoundation.org/

Appendix A: Transcripts of real-life queries submitted to LC

1 Ladies and Gentlemen, please send me by e-mail or give me access to picture file(s) of: Frontpage of (any) US newspaper (i.e. Stars and Stripes) heading the meeting of US and Soviet troops on April 25th, 1945 at Elbe River near Torgau / Germany. In case you cannot fulfill my request, I kindly ask you to forward it to a library/archive which is able to do so! Thank you for care and efforts.

2 I have found letters from George Hoadley to his father when he was the Washington correspondent of the United States Gazette. These letters are from 1806-1808. They are fascinating impressions of his about the beginning of our Republic. I would like to find out more about the Gazette & his writings. He mentions Jefferson, Madison etc. I feel these have great historical significance. Thank you for your help.

3 I'm doing a research paper on Nat Turner's rebellion. Unfortunately, I'm in Oregon and most of the Virginian newspapers from 1831, are not. I was wondering how to access the information in the newspapers from Virginia during 1831.

4 I am looking for a newspaper published on or after October 21, 1944. I have a clipping with my grandfather in it, but no reference to the publication from which it came. The caption below the picture reads "Atlantian throws ACK-ACK at Japs" and mentions his name, JACK W. MEADOW. According to the caption, the picture was taken on October 21, 1944 after a battle at "SUNGSHAN MOUNTAIN". On the opposite site of the clipping is an unrelated article by C.E. GREGORY. His article's title is "Increased Cost Of State Farm in Line with Program". Though it does not mention which state it comes from, the focus indicates it is a local/georgia newspaper. How do you recommend I continue my search for this newspaper?

5 I'm a journalism college professor. I'd like to know if La Patria (a New Orleans-based newspaper) was the first Spanish-language daily in the U.S. Who was the publisher/editor? Also, was La Gaceta the second Spaish language newspaper in U.S. (possibly Texas) both newspapers were around in the 1800s. Which came first. Even our testbook is confusing.

6 Do you have the Nashville Banner in any form prior to 1948? (I am trying to locate the original version of a pair of articles I have copies of. About all I know is title and name of the author and that they appeared on two successive Sundays. I presume it would be necessary to search actual copies, as there seems to be no index).

7 I am looking for an article in that ran in the Denver Post on 04/26/07. It was written by Hugh O'Neill and covered the events around the death of Alfred Packer (infamous Colorado cannibal). I would like to know how to receive a copy of the masthead, showing the date, the headline of the article and the body of the article—preferably something like a jpeg or something legible we can read and film with a camera to use for a show for National Geographic Channel. Please let me know what to do next as I saw your online search for publications that old are under construction.

8 I would like to know if you have copies of a newspaper that was only published 5 months in Butler, Pennsylvania from Dec. 1879-April 1880? It is called "The Prospect Leader", and also the Petrolium Record published about the same time in PA.

9 General Inquiry:
I am trying to locate the text of two articles published in the Brooklyn Daily Advertiser, 1850. Both are by Walt Whitman. They are titled "Works of Beauty and Talent-The New Art Union of Brooklyn", 4 April 1850 and "Brooklyn Art Union-Walter Libbey-A Hint or Two on the Philosophy of Painting". 21 December 1850.
10
10 I am trying to locate a German/American newspaper which my great uncle published in Lafayette Indiana in the 1870s, 1880s, and 1890s called Der Deutsche Amerikaner. It is very important that I find some copies. Can you help me?

11 I am seeking information about the completeness of the LOC's microfiled copy (newspaper microfilm 2679) of the serial Tiesa (Brooklyn, N.Y.) LC control # sn 86071641. The publisher, the Association of Lithuanian Workers, is planning to merge with another organization and I am trying to determine if a complete collection of this serial already exists on microfilm. I am seeking to offer assistance to the Association of Lithuanian Workers to preserve their organizational archives before they merge.

12 I am attempting to ascertain whether a “petition“ published on page 2 of the Piladelphia Packet and Daily on Friday, December 8, 1786 was also published in other contemporaty papers. From an unnamed correspondent, it purports to be from single women in state capitals who want the Congress to meet there so they can meet eligible bachelors. It is titled “The PETITION of young Ladies of Porsfmoutyh, Bofton, Newport, New London, Amboy, News-Cafile, Williamfburg, Wilmington, Charles and Savannah“. I believe it to be a paradoy, probably by a northern writer who does not know that the capital of VA had been moved to Richmond. The petition sounds as though it could have been written by Benjamin Franklin (who lived in Philly), but it might also be a reprint from another publication of the day.

13 Can you please tell me when Metz Lochard starting working for the Chicago Defender. What year his name first appeared in the paper and what he did or any other information you may have about him. Also is there any way you can help or persuade the Chicago Historical Society to film the copies of the Metropolitan News, they are starting to fall apart and I believe they have the only copies around.

14 To the Digital Reference Team: I am looking for old newspaper articles concerning Indian treaties of 1854-55 of Washington territory, I would like to read from papers popular with Congress and the industrial elite, articles from ..., NYNY, Philly, ... Could you help me find the online archive? I am writing a speach for a commemoration dinner for the 150th anniversary of the Treaty of Neah Bay 1855.

15 I understand that you have a file of a newspaper, the Washington, D.C. “Union“? (LC Control No. Sn 82006534/ Call No. Newspaper 7432). I am interested in knowing whether Vol. 13, no. 1 (April 18, 1857) through Vol. 14, no. 304 (April 10, 1859) of the “Union“ is available on microfilm and can be borrowed on the inter-library loan through my local public library (Bloomfield Township, Michigan Public Library). If it is on film and can be borrowed, who at LOC should my library contact to arrange this? If your file of the Washington “Union“ is not on microfilm, (1) do you know of a repository elsewhere that has a microfilm copy or (2) can you tell me how I might go about having a film made of your paper file?

16 To the Digital Reference Team: Clarence Brigham’s History and Bibliography of American Newspapers, 1690-1820 (1947) says the LOC owns issues of a Washington-Kentucky newspaper called The Miror running from Oct. 21, 1797-Sept. 4, 18, 1799. The LOC Online Catalog does not show that you own any copies of that newspaper. Is the Online Catalog wrong? If you once had them, but no longer do, do you know what became of those issues?

17 The late Hunter S. Thompson once made reference to Yellowboy Willis. I think he might have associated with yellow journalism. Any clues?

18 To the Digital Reference Team: I’m trying to find an article that would have appeared in a newspaper in Cumberland, MD from May of 1928. Does the LOC have archives from either the Cumberland Times News or the Cumberland News? I’m contacting the paper directly as well, but not sure what they will have.

19 My brother was murdered in Redondo Beach in 1985. I was wondering if any of your newspapers have any articles in them, about this murder? Thanks so much.

20 Does the LC hold original newspapers in its collection, and if so, are older volumes ever discarded? I am interested in finding original copies Swedish-American newspapers from my hometown, particularly SVEA (1897-1965 Worcester, MA) or SKANDINAVIA (1881-1918 Worcester, MA). I am a Swedish-American historian/author with a rather extensive personal collection and I have had no luck in finding original copies for my studies. If the LC does discard materials over time, is there a listing of such materials? I am not sure of the process or if there is a policy rearding such material.

21 My grandmother came to Ellis Island from Genoa, Italy on the SS San Giovanni. The ship arrived in Ellis Island on Feb 2, 1921. On the way to America this ship sailed around a major storm in the Atlantic in late January 1921. I would like to have a copy of a NY City newspaper article about the storm. I know such an article exists. The article I seek named the ships not heard from since the storm began. How can I find this article?

22 Dear Ms./ Mr. Librarian. I’m writing to you to request the newspaper research. I’m a researcher from South Korea. I want to read the newspapers of the historical period, 1945 (the New York Times and the Washington Post). I’m looking for the bound type news papers not microfilm if it’s available. Furthermore, one

of my colleagues wants to shoot film some articles in that newspapers. The newspapers I want to see are the following. The New York Times, June 30, 1945 Article: russians Captured with Nazi Riot at Fort Dix pg. 1.2. The New York Times , July 1, 1945 Article: U.S. halts returns of 150 to Russia pg. 4.1. The Washington Post, July 1, 1945 Article: Reds unwilling return home halted at pier . I want to read these three articles in the library. I want to visit the library with my two colleagues in the next Monday or Tuesday (Nov. 7 or 8 in the morning). Could you let me know if it's possible or not to film the newspaper's article.

23 Dear Sir/Madam: I am writing to you from Moscow, Russia. I am a private person, I'm 29, and looking for my relative's history as we have lost any information about him long ago. I have written to the New York Times and they recommended me to turn to Archive Libraries with regard to my request. I have relatives who live in USA at present and the eldest of them have witnessed themselves the articles regarding our relative in New York Times. Unfortunately I do not have exact dates and years of this publication. The article was about death and funeral of the Russian Lieutenant Shmidt, years of publication may be from 1905 to 1917. I would be extremely grafeful to you if you could answer me if it's possible to receive a copy of the article and what is even more important photo attached to it, or advise me where I could go to search next. Or should I make an official request or come personally? (I quite realize that no one would be happy to sit down and spare his time searching old date bases for many years for an absolutely unknown person). But I've heard a lot about how you preserve your history and I am writing to you with the hope of getting any help. I would appreciate for any answer (even negative). Thank you in advance. Remaining sincerely yours, ...

24 The American Memory Team: Hello! I am doing some historical research, and I am hoping that you would please look in a reference book that would list newspapers that were being published in 1810. I am only interested in newspapers that were being published in Norfolk, Virginia in 1810, and also in the Mississippi Valley area in 1810 (including present-day Louisiana, Mississippi, Alabama, Tennessee, Arkansas, Missouri and Illinois). Do you know of a reference book that would note where copies or microfilm of these newspapers might be located?

25 I have been trying to find our where I can obtain a copy of LOC Microfilm # 2032, Royal Danish American Gaztette 1770-1779 for about two months now. I keep being told that I need to readdress my request. I think I have spoken to at least 20 people long distance on telefone. I am told that the master microfilm is not available to make duplicates for sale. When I asked where you would go to have a new copy made if your copy was damaged, they hung up on me. I wish to obtain a copy for our local library here in St. Croix. Do I need to go to the Director of the Library of Congress?

26 To the American Memory Team: My family is having a 90th birthday party for our dad in October. One segment of celebration program is history/trivia. I already have the "birthday times" indicating some facts from 1915, i.e. the cost of bread and milk. I also searched the web for historical events in 1915. I am aware that the Lusitana was attacked 1915 and subsequently caused the US to enter WWI. Also the film "Birth of a Nation" was relased in 1915. However, I wanted the headlines from the newspaper on the day he was born 10-20-15. Also I wanted to contrast main stream newspaper headlines, i.e. Wasington Post vs African-American headlines, i.e. the Afro on that date. Also as an alternative I would like articles of interest during the entire year of 1915. The additional facts would be used in completing the trivia presentation document. Where should I go at the library to find the newspapers from 1915? Is this resource available on Saturdays? Also are there books that would give some of the major events of 1915? Maybe periodicals are available. I don't know how long Time or Newsweek have been in existence. Are there any historical websites that are free?

27 Chat Session Transcript: I'm trying to find anything I can about a man who walked around the border of the U.S. in 1910 named John (Jack) Albert Krohn. I have one book about him, but can't seem to find literally anything else.

28 I am researching a book on the Nancy Harts, a female Civil War militia from LaGrange, Georgia. I understand that it was the only female militia commissioned by either side. To discern whether this is true, I am looking for an official record of the commission. Can I get it via the Library of Congress. Additionally, I am looking for any extant copies of the LaGrange Reporter newspaper, which I understand is virtually impossible to find. Do you have any copies? Thank you so much for this priceless service.

29 I am researching the publication history of John Lawson's "New Voyage to Carolina" published in London in 1709 in 4 parts in John Stevens' "New Collection of Voyages and Travels". This is an important early work from colonial North Carolina. The book was reportedly being advertised for sale as late as 1723, according to R.M. Wiles "Serial Publication in England before 1750" (Cambridge University Press, 1957) footnote 2 on

page 87. Wiles cites “St. Jame’s Evening Post“ of London issue # 1043 for 25 January 1723 as having the advertisement. I have not been able to locate a copy of this particular issue in any microfilm or hard copy reported or catalogued. Do you have any suggestions on where or how to search for a copy of this newspaper?

30 Hello. Does the Library of Congress own the newspaper:Courrier de la France et des colonies, published in Philadelphia, PA Oct. 15, 1795 – March 14, 1796 by Louis F.R.A. Gatereau and printed by Moreau de Saint-Mery? If so, does LC have it as the original newspaper or is it a microform or photocopy version? RLIN lists LC as owning a copy but I was not able to find an entry for it in the regular LC online database, the LC Serial and Govenment Publications Division’s Eighteenth Century American Newspapers in the Library of Congress database, the LC Serial and Government Publications Division’s Commonly used newspapers on microfilm database, nor in OCLC with an LC holdings attached. I am asking for confirmation of LC’s holdings because the Boston Athenaeum will be celebrating their bicentennial soon and we are planning on showing this newspaper in an exhibition (& catalog) which highlights the collecting we have done over the past 200 years. The Athenaeum has always believed that we own the only known actual physical copy of this newspaper and I want to confirm if this is in fact true before the exhibition goes up. If LC also owns a physical copy of the newspaper it would be very exiting and we would want to record this in our records.

31 Dear Sir/Madam, I am helping a colleague .... in his preparations of a manuscript concerning the visit to the United States by Henry James in 1904-1905, and his subsequent travel book, The American Scene. .... wishes to include a caricature of the novelist entitled “Henry James’ First Appearance“, which appeared in the Seattle Post-Intelligencer on Sunday 29 January 1905, on p. 8, under the heading “Fame and Fun“. It first appeared in Philadelphia Press (I presume shortly before, but I don’s have a date), which is no longer extant. ... has a photocopy of this image, which he obtained from visiting the Library of Congress, but needs it in some format, possibly microfiche or microfilm , from which a version suitable for publication could be scanned. I have searched your on-line catalogue but am unable ti find it. I am also trying to locate a good image of another caricature called “f I were Henry James“ which appeared in metropolitan magazine (21:364 December 1904).Any help you could provide regarding either of these images would be greatly appreciated.

32 Dear Sir/Madam, By way of introduction, I am writing to you from Granada Television in the United Kingdom. We are currently in the latter stages of producing a technology-based documentary series, ‘Warplane‘ for PBS in the USA, scheduled for transmission worldwide in 2006. The co-production partners are WNET 13 New York, National Geographic and Channel 5 UK. Warplane traces the evolution of military aircraft from the Wright Brothers to the present day, with emphasis on evolving technology. Filmed with the active co-operation of the Pentagon and NATO, the series comprises four hour long episodes in which we trace the evolution of the warplane from the first Wright Brothers military commission in 1908 through to modern times, the B-2, the F-22, the Eurofighter, and beyond. We are keen to source images from the New York Times and The Washington Post announcing the death of Chuck Yeager. I understand most newspapers of the time are held on microfilm. I wonder if you know of an archive that stores the original copies.

33 I am producing a film on the Klan, civil rights and lynching. I need film (transferred to mini dv cassettes) and stills from the period of 1865 – 1965. How do I research this? I need: film clips and stills of the Klan, segregation, Dr. Martin Luther King, Jr., civil rights marches and protests, stills of any lynchings, old newspapers discussing lynchings. Once I locate the iems to order, what is the charge for (1) still and (1) 15-60 seconds film clip and (1) newspaper duplication? These items will need to be in the Public Domain.

34 To the American Memory Team: I am searching for The Daily American Star. I know it is a newspaper published from September 1847 through March 1848. I will appreciate any information you can provice, also if I can get some copies how can I do it.

35 What was the first daily newspaper in American history? I am really looking for print resourse with this information in it? Can you direct me to one?

36 Hello, my name is .... and I am a research archivist with the Maryland State Archives. I am looking for information regarding the early american newspaper, “The Mail, or, Claypoole’s Daily Advertiser“. It was published in Philidelphia Pennsylvania between 1791 and 1793. I am looking for any information about the politics of the newspaper, especially as it relates to slavery and abolition. Any more information you could provide me would be invaluable.

Appendix B: Retrieval results at page level and article level

	query	page-level hits p	article-level hits a	relevant hits at page level pa	relevant hits among first 5 page-level hits	relevant hits among first 10 page-level hits
1	+potomac ++frozen	19	3	3	1	2
2	+potomac +river +frozen	9	2	2	2	2
3	+potomac +steamship	161	4	4	1	2
4	+potomac +hampton	21	1	1	0	0
5	+japanese +immigrants	15	4	4	1	2
6	+open +door +china	174	28	28	2	4
7	+open +door +china +roosevelt	62	9	9	2	3
8	+wilson +taft	87	16	16	1	2
9	+roosevelt +taft	173	105	99	5	10
10	+roosevelt +wilhelm	19	6	6	2	4
11	+hay +root	43	20	22	4	6
12	+roosevelt +russia +japan	213	79	90	2	7
13	+refrigerator +fruit	57	7	7	4	5
14	+railway +fares +roosevelt	5	1	1	1	1
15	+rockefeller +carnegie	14	9	9	4	6
16	+carnegie +gospel +wealth	5	2	2	2	2
17	+gospel +wealth	12	4	4	3	4
18	+carnegie +trade +union	21	0	0	0	0
19	+carnegie +pacific +union	23	0	0	0	0
20	+morgan +pacific +union	73	6	6	0	1
21	+morgan +railway	108	17	17	2	2
22	+morgan +McKinley	16	5	5	2	4
23	+morgan +cab +accident	8	1	1	1	1
24	+taft +philippines	82	56	59	5	10
25	+panama +canal +immigrants	8	2	2	0	2
26	+panama +canal +railroad	217	56	56	4	9
27	+big +stick +america	59	14	14	3	3
28	+russia +japan	376	289	286	5	10
29	+george +single +tax	50	1	1	0	1
30	+firemen +pension	55	8	7	2	4
31	+firemen +pension +rights	4	1	1	1	1
32	+smallpox +remedy	25	6	6	2	4
33	+smallpox +cure	29	5	5	1	2
34	+smallpox +drug	14	3	2	0	1
35	+beef +trust	163	87	84	4	7
36	+beef +trust +roosevelt	79	16	15	2	3
37	+beef +trust +gorki	1	0	0	0	0
38	+beef +trust +william	130	4	4	0	1
39	+beef +trust +wilhelm	7	0	0	0	0
40	+suicide +student	31	13	12	3	5
41	+funeral +rothschild	13	0	0	0	0

42	+funeral +miller	159	34	29	2	5
43	+storm +atlantic	144	21	21	1	2
44	+indian +treaties	10	4	4	2	4
45	+henry +james	788	382	319	4	9
46	+atlantic +storm	144	21	21	1	2
47	+ellis +island	223	23	23	3	6
48	+morocco +roosevelt	38	5	5	0	1
49	+entente +cordiale	3	3	3	3	3
50	+balfour +resign	8	6	6	4	6
51	+protective +tariff	15	12	11	5	7
52	+department +agriculture	236	178	172	5	9
53	+department +justice	665	216	208	3	5
54	+morgan +h. +beach	131	28	29	2	4
55	+roosevelt +europe	228	83	79	4	6
56	+new +york +church	1125	423	328	4	8
57	+southeast +washington	596	203	192	5	10
58	+memorial +services +miller	53	13	12	2	3
59	+memorial +services +jones	58	6	6	0	1
60	+marriage +miller	154	41	39	4	6
61	+marriage +jones	179	39	37	2	3
62	+gulf +stream +activity	7	3	3	3	3
63	+fant +buried	2	1	1	1	1
64	+yacht +clubhouse	9	4	4	3	4
65	+secret +service +cotton	95	21	25	4	7
66	+jurors +coroner +murdered	3	2	2	2	2
67	+iowa +china	13	3	3	1	2
68	+prominent +clerk	240	40	39	3	5
69	+social +organisation +louisville	1	0	0	0	0
70	+social +organization +louisville	32	2	2	1	2
71	+afghanistan +troops +mission	2	1	1	1	1
72	+memorial +day +parade	27	8	8	2	4
73	+ambassador +mexiko	0	0	0	0	0
74	+ambassador +mexico	85	42	42	5	9
75	+fire +life	705	174	176	3	7
76	+law +snow	182	63	67	5	10
77	+poor +boy +leading	105	5	5	2	2
78	+court +child	450	101	99	4	8
79	+norway +crisis +sweden	13	8	8	5	8
80	+roosevelt +texas +arrived	67	5	5	1	1
81	+suicide +student	31	13	12	3	5
82	+penalty +priest	11	6	6	4	5
83	+john +paul +jones	262	77	75	2	6