IETF NNTP Working Group
Internet Draft

N. Ballou, Microsoft
B. Hernacki, Netscape
S. Waters, Microsoft

Document: draft-ietf-nntpext-srch-00.txt

January, 1998

NNTP Full-text Search Extension

Status of this Memo

1. Abstract

This document describes a set of enhancements to the Network News Transport Protocol [NNTP-977] that allows full-text searching of news articles in multiple newsgroups. The proposed SEARCH command supports functionality similar to the [IMAP4] SEARCH command, minus user specific search keys (i.e., ANSWERED, DRAFT, FLAGGED, KEYWORD, NEW, OLD, RECENT, SEEN) and minus search keys based on headers that do not exist in news (i.e., CC, BCC, TO).

The availability of the extensions described here will be advertised by the server using the extension negotiation-mechanism described in the new NNTP protocol specification currently being developed [NNTP-NEW].

2. Conventions used in this document

In examples, "C:" and "S:" indicate lines sent by the client and server respectively.

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in [RFC-2119].

3. Introduction

The NNTP SEARCH command is sent from the client to the server to specify and initiate a full-text search on articles in one or more newsgroups. The NNTP SEARCH command is similar to the [IMAP4] SEARCH command, with user property and mail-specific header search keys not present in NNTP SEARCH. The results of an NNTP Search is OVER data as specified in [NNTP-NEW] for each article that satisfies the search criteria.

In addition, the PAT command is extended so that it can be used to full-text search articles within a single newsgroup. Both the headers and the body of the articles are searched.

3.1. New and Enhanced NNTP Commands

There are four new NNTP commands: two new options to the existing LIST command, and enhancements to one existing command.

* SEARCH
* LIST SRCHFIELDS
* LIST SEARCHABLE
* PAT

The SEARCH command runs a one-time search, returning overview-like data.

The LIST SRCHFIELDS command returns the fields that the server allows in full-text searches.

The LIST SEARCHABLE command allows the client to determine which newsgroups are full-text searchable.

The PAT command allows the pseudo-header ":TEXT". This specifies a full-text (headers and body) search of the articles in a single newsgroup.

4. Use of NNTP Extension Mechanism

The NNTP extension mechanism allows a server to describe its

capabilities. The following extensions are used to describe the capabilities described in this document.

4.1. SRCH Extension

The SRCH extension means that the server supports the following commands: SEARCH, LIST SEARCHABLE, LIST SRCHFIELDS.

4.2. PATTEXT Extension

The PATTEXT extension means that the server supports the :TEXT header in the PAT command, as described by this document.

5. SEARCH Command

```
Arguments:   optional newsgroup specification
             searching criteria (one or more)

Responses:   224 overview information follows
             412 no news group selected
             462 error performing search
             463 too many hits
             480 authentication required
             501 command syntax error
             502 no permission
```

The SEARCH command searches the newsgroups for articles that match the given searching criteria. Searching criteria consist of one or more search keys. If there are articles that match the search criteria, the server responds with code 224 and returns OVER data for each matching article in a similar format as described in [NNTP-NEW] with one exception. The one change from [NNTP-NEW] OVER format is to change the article number field to a format that supports searches over multiple newsgroups. The article ID field for SEARCH OVER data will use the format newsgroup:art-ID rather than just an article number as defined in [NNTP-NEW] (note: this is the same format used by the Xref header).

A response of 501 indicates a syntax error in the search criteria. A response of 502 indicates that the user does not have permission to search one or more of the specified newsgroups. If the search criteria did not specify a newsgroup, and there is no current newsgroup (i.e., set using the NNTP GROUP command), then the server returns the error code 412, indicating that no newsgroup has been specified. A response of 462 indicates that the server encountered an error when processing the search. Some

implementations may wish to limit the maximum number of articles that can match a search. A response of 463 indicates that the number of hits has been exceeded. This response may also be used to indicate that a search request is not being performed because it is anticipated to produce too many matches. An example would be searching for a single character.

When multiple keys are specified, the result is the intersection (AND function) of all the messages that match those keys. For example, the criteria FROM "SMITH" SINCE 1-Feb-1994 refers to all articles from Smith that were placed in the newsgroup since February 1, 1994. A search key may also be a parenthesized list of one or more search keys (e.g. for use with the OR and NOT keys).

Server implementations MAY exclude [MIME-1] body parts with terminal content types other than TEXT and MESSAGE from consideration in SEARCH matching.

The optional newsgroup specification consists of the word _IN_ followed by either a wildcard character _*_ - indicating a search over all newsgroups - or a list of newsgroup names separated by a comma. A newsgroup name can end with the wildcard string _.*_ indicating a search over a sub-hierarchy of the newsgroup name space. If no newsgroup specification is given, the search is over the current newsgroup. If there is no current newsgroup, the server returns the 412 error code.

The ON, BEFORE, and SINCE search criteria use the same date as used in the NNTP NEWNEWS command in [NNTP-NEW] - the date the article arrived on the server. A server indicates support for the ON, BEFORE, and SINCE search criteria by listing :Date in the LIST SRCHFIELDS response.

The defined search keys are as follows. Refer to the Formal Syntax section for the precise syntactic definitions of the arguments.

<message range>	Articles with article numbers corresponding to the specified range.
ALL	All Articles in the current newsgroup; the default initial key for ANDing.
BEFORE <date>	Articles whose server arrival date is earlier than the specified date.
BODY <string>	Articles that contain the specified string in the body of the message.
FROM <string>	Articles that contain the specified string in the article structure's FROM field.

HEADER <field-name> <string>
Articles that have a header with the specified field-name (as defined in [RFC-822]) and that contains the specified string in the [RFC-822] field-body. If the <string> is the empty string (""), then a match implies that the specified header does not exist.

LARGER <n> Articles with a size larger than the specified number of octets.

NOT <search-key> Articles that do not match the specified search key.

ON <date> Articles whose server arrival date is within the specified date.

OR <search-key1> <search-key2> Articles that match either search key.

SENTBEFORE <date>
Articles whose [RFC-822] Date: header is earlier than the specified date.

SENTON <date> Articles whose [RFC-822] Date: header is within the specified date.

SENTSINCE <date>
Articles whose [RFC-822] Date: header is within or later than the specified date.

SINCE <date> Articles whose server arrival date is within or later than the specified date.

SMALLER <n> Articles with a size smaller than the specified number of octets.

SUBJECT <string>
Articles that contain the specified string in the envelope structure's SUBJECT field.

TEXT <string> Articles that contain the specified string in the header or body of the message.

Example:
C: SEARCH FROM "Smith" SINCE 1-Feb-1994
S: 224 overview information follows
S: comp.object:573 \t RE: object-oriented langs \t \
"John Smith" <JSmith@xyz.com> \t Sun, 03 Nov 1996 \
14:25:05 -0800 \t <01cbc9d5f3c70$eab9a2cd@xyz.com> \

```
          \t 4080 \t 33
       S: .
```


Note: each field in OVER response is separated by a tab - shown as a \t in the example above.

5.1.1. Search Formal Syntax

The search query syntax is derived from the search syntax defined for the IMAP4 protocol. It is somewhat different because of the way international character sets need to be encoded. The following syntax specification uses the augmented Backus-Naur Form (BNF) as described in [ABNF].

Except as noted otherwise, all alphabetic characters are case-insensitive. The use of upper or lower case characters to define token strings is for editorial clarity only. Implementations MUST accept these strings in a case-insensitive fashion.

```
  astring       ::= atom / string

  atom          ::= 1*ATOM_CHAR

  ATOM_CHAR     ::= <any CHAR except atom_specials>

  atom_specials ::= "," / "(" / ")" / SPACE / CTL / "*" /
                    quoted_specials

  CHAR          ::= <any ASCII character except NUL, 0x01 - 0x7f>

  CTL           ::= <any ASCII control character and DEL, 0x00 -
                    0x1f, 0x7f>

  date          ::= date_text / <"> date_text <">

  date_day      ::= 1*2digit
                    ;; Day of month

  date_month    ::= "Jan" / "Feb" / "Mar" / "Apr" / "May" / "Jun"
                    / "Jul" / "Aug" / "Sep" / "Oct" / "Nov" /
                    "Dec"

  date_text     ::= date_day "-" date_month "-" date_year

  date_year     ::= 4digit

  digit         ::= "0" / digit_nz

  digit_nz      ::= "1" / "2" / "3" / "4" / "5" / "6" / "7" / "8"
```

```
                    / "9"


     header_fld_name ::= sstring
INTERNET-DRAFT      NNTP Full-text Search Extension      January 1998


     mstring       ::= A MIME encoded string surrounded by double
                       quotes

     newsgroup     ::= atom [ ".*"]

     newsgroups    ::= "*" / newsgroup_list

     newsgroup_list ::= newsgroup [ "," newsgroup_list]

     number        ::= 1*digit
                   ;; Unsigned 32-bit integer
                   ;; (0 <= n < 4,294,967,296)

     nz_number     ::= digit_nz *digit
                   ;; Non-zero unsigned 32-bit integer
                   ;; (0 < n < 4,294,967,296)

     QUOTED_CHAR   ::= <any TEXT_CHAR except quoted_specials> / "\"
                       quoted_specials

     quoted_specials ::= <"> / "\"

     range         ::= nz_number / nz_number "-" [ nz_number ]
                   ;; Identifies a range of Articles.

     search        ::= "SEARCH" SPACE
                       ["IN" SPACE newsgroups SPACE]
                       1#search_key

     search_key    ::= "ALL" / "BODY" SPACE sstring / "FROM" SPACE
                       sstring / "ON" SPACE date / "SINCE" SPACE date
                       / "BEFORE" SPACE date / "SUBJECT" SPACE
                       sstring / "TEXT" SPACE sstring / "HEADER"
                       SPACE header_fld_name SPACE sstring / "LARGER"
                       SPACE number / "NOT" SPACE search_key / "OR"
                       SPACE search_key SPACE search_key /
                       "SENTBEFORE" SPACE date / "SENTON" SPACE date
                       / "SENTSINCE" SPACE date / "SMALLER" SPACE
                       number / range / "(" 1#search_key ")"

     SPACE         ::= 1*<ASCII SP, space, 0x20>

     sstring       ::= astring / mstring

     string        ::= <"> *QUOTED_CHAR <">
```

```
  TEXT_CHAR     ::= <any CHAR except CR and LF>
```

5.2. LIST SRCHFIELDS Command

```
   Arguments: none
```



```
   Responses: 224 data follows

   The LIST SRCHFIELDS command returns a list of which fields can be
   specified in full-text search queries on the server. The response
   is a list of searchable fields, one per line. A _._ on its own
   line terminates the list.  The fields are either newsgroup
   headers, or non-header fields supported by the query syntax.

   The three currently defined non-header fields are _:Body_,
   _:Text_, and _:Date_. _:Text_ means all the searchable text in the
   article, and indicates that the _TEXT_ keyword is supported in the
   search query language. _:Body_ means the body of the article,
   excluding the headers, and indicates that the _BODY_ keyword is
   supported in the search query language. _:Date_ means the date at
   which an article arrived on a server - similar to the date used in
   the NNTP NEWNEWS command - and indicates that the _ON_, _SINCE_,
   and _BEFORE_ keywords are supported in the search query language.

   The _TEXT_ and _BODY_ search query fields are optional, but the
   server must indicate whether they are supported or not in the LIST
   SRCHFIELDS response.

     Example: C: LIST SRCHFIELDS
              S: 224 Data follows.
              S: From
              S: Date
              S: Subject
              S: :Text
              S: .
```

5.3. LIST SEARCHABLE Command

```
   Arguments: none

   Responses: 224 Data Follows

   The LIST SEARCHABLE command returns a list of strings that define
   which new groups are being indexed by the news server and are thus
   available for searching. In addition, the character sets allowed
   for each group is returned.
```

When there are newsgroups indexed it will return 224, followed by each portion of the tree that is indexed. If all groups are indexed, a line with "*" is returned. If only some parts of the newsgroup hierarchy are indexed, they are identified in the form <indexed-hierarchy>.*. Clients should not assume that these will always be top level hierarchies. A "." on its own line terminates the list.


```
Example: C: LIST SEARCHABLE
         S: 224 Data follows.
         S: alt.*
         S: comp.lang.*
         S: mcom.*
         S: .
```

5.4. PAT Command Enhancement

Arguments: header range|<message-id> [pat [pat...]]

Responses: <same as PAT - see [NNTP-NEW]>

The PAT command is enhanced in a simple way: The new value _:TEXT_ will be supported as a header when invoking the command. The :TEXT header requests a full-text search of the body and all headers of the specified articles. Other than adding a new header name, the PAT command arguments are the same as specified in [NNTP-NEW].

If :TEXT isn't specified as the header, the response is the same as it always has been for PAT, with each result line containing the article number and the value of the header that matched the pattern.

If the :TEXT header is specified, the constant string _TEXT_ is returned in place of the value of the header that matched the pattern.

```
Example: C: PAT :TEXT 1000-2000 searchtext
         S: 221 Header follows
         S: 1021 TEXT
         S: 1024 TEXT
         S:.
```

6. Security Considerations

The search commands must be implemented in a way that does not allow access to articles in newsgroups that a client is otherwise

restricted from reading due to access control rules.

7. References

[ABNF], DRUMS working group, Dave Crocker Editor, _Augmented BNF for Syntax Specifications: ABNF_, draft-drums-abnf-02.txt (work in progress), Internet Mail Consortium, April 1997

[IMAP4] IMAP4 INTERNET MESSAGE ACCESS PROTOCOL - VERSION 4rev1. M Crispin, Request for Comment (RFC) 2060, December 1994

[MIME-1] Borenstein N., and N. Freed, MIME (Multipurpose Internet Mail Extensions) Part One: Format of Internet Message Bodies, Request for Comment (RFC) 2045, December 1996.

[NNTP-977] Network News Transfer Protocol. B. Kantor, Phil Lapsley, Request for Comment (RFC) 977, February 1986.

[NNTP-NEW] Network News Transfer Protocol. S. Barber INTERNET DRAFT, draft-ietf-nntpext-base-02.txt, September 1997.

[RFC-2119], Bradner, S, _Key words for use in RFCs to Indicate Requirement Levels_, RFC 2119, Harvard University, March 1997

8. Acknowledgments

TBD

9. Author's Addresses

Nathaniel Ballou
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399
Phone: 425-703-0574
Email: natba@microsoft.com

Brian Hernacki
Netscape Communications
501 E. Middlefield Rd.
Mountain View, CA 94043-4042
Phone: 650-937-6738
Email: bhern@netscape.com

Stephen Waters
Microsoft Corporation
One Microsoft Way
Redmond, WA 98052-6399
Phone: 425-703-4972

Email: swater@microsoft.com