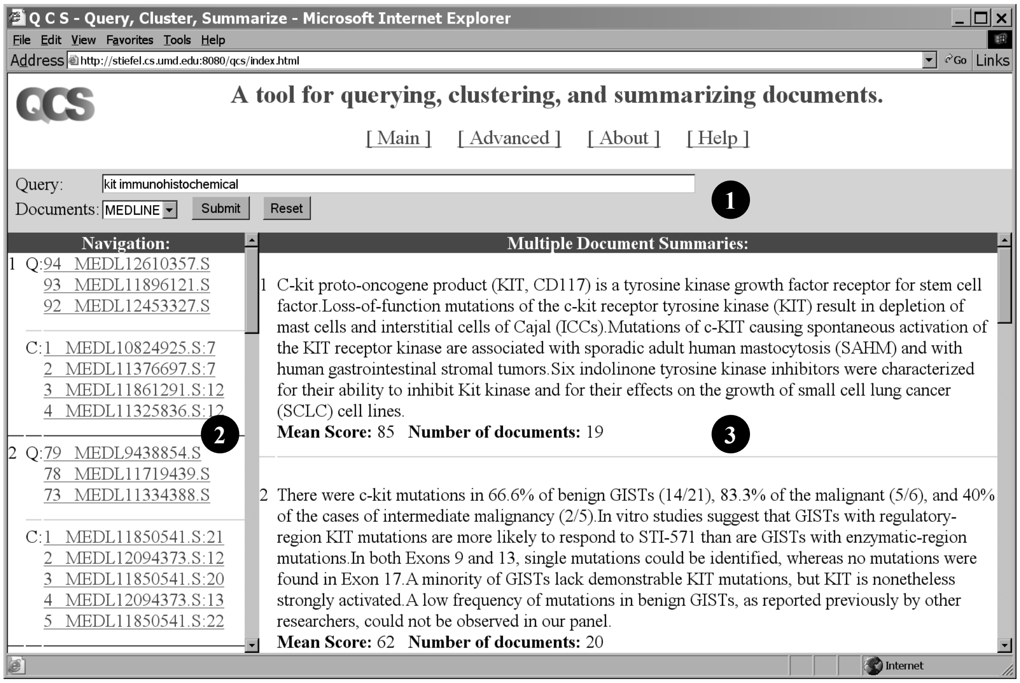
Clustering and Summarizing Medline Abstracts

Dunlavy, Daniel M.*[*1], Conroy, John M.[2], O'Leary, Timothy J.[3], O'Leary, Dianne P.[1]
[1]University of Maryland, College Park, MD, USA; [2]Center for Computing Sciences, Bowie, MD, USA; [3]Armed Forces Institute of Pathology, Rockville, MD, USA

The amount of reference material available to medical professionals is overwhelming, and its usefulness depends not only on timely delivery but also on information triage. The MEDLINE system is invaluable for extracting documents relevant to a query and presenting them in a rank-ordered list. When a query returns a large number of documents, however, it is often impractical to search this list for desired documents.

In previous work, we have developed a system to extract documents relevant to a query, cluster these documents by subject, and return a summary of the documents in each cluster augmented by the list of documents in that cluster. This Query-Cluster-Summarize (QCS) system was designed with newswire documents in mind. We have tested it on MEDLINE documents related to gastrointestinal stromal tumors (GIST), and the results are encouraging.

A screenshot of the QCS system is shown below for the query "kit immunohistochemistry". The interface to the QCS system includes 1) an *input* section for the query and choice of document set, 2) a *navigation* section with links to clustered documents (Q: top documents retrieved for the query and their scores, C: documents from which the summary sentences were drawn and the sentence indices), and 3) an *output viewing* section, which here contains the default output of multiple document summaries for the first two topic clusters.


Q C S - Query, Cluster, Summarize - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Address http://stiefel.cs.umd.edu:8080/qcs/index.html
Go Links
QCS
[Main] [Advanced] [About] [Help]
Submit
Reset
1
2
3
Internet

A tool for querying, clustering, and summarizing documents.

Query: kit immunohistochemical
Documents: MEDLINE

Navigation:

1 Q:94 MEDL12610357.S
93 MEDL11896121.S
92 MEDL12453327.S

C:1 MEDL10824925.S:7
2 MEDL11376697.S:7
3 MEDL11861291.S:12
4 MEDL11325836.S:12

2 Q:79 MEDL9438854.S
78 MEDL11719439.S
73 MEDL11334388.S

C:1 MEDL11850541.S:21
2 MEDL12094373.S:12
3 MEDL11850541.S:20
4 MEDL12094373.S:13
5 MEDL11850541.S:22

Multiple Document Summaries:

1 C-kit proto-oncogene product (KIT, CD117) is a tyrosine kinase growth factor receptor for stem cell factor.Loss-of-function mutations of the c-kit receptor tyrosine kinase (KIT) result in depletion of mast cells and interstitial cells of Cajal (ICCs).Mutations of c-KIT causing spontaneous activation of the KIT receptor kinase are associated with sporadic adult human mastocytosis (SAHM) and with human gastrointestinal stromal tumors.Six indolinone tyrosine kinase inhibitors were characterized for their ability to inhibit Kit kinase and for their effects on the growth of small cell lung cancer (SCLC) cell lines.
Mean Score: 85 **Number of documents:** 19

2 There were c-kit mutations in 66.6% of benign GISTs (14/21), 83.3% of the malignant (5/6), and 40% of the cases of intermediate malignancy (2/5).In vitro studies suggest that GISTs with regulatory-region KIT mutations are more likely to respond to STI-571 than are GISTs with enzymatic-region mutations.In both Exons 9 and 13, single mutations could be identified, whereas no mutations were found in Exon 17.A minority of GISTs lack demonstrable KIT mutations, but KIT is nonetheless strongly activated.A low frequency of mutations in benign GISTs, as reported previously by other researchers, could not be observed in our panel.
Mean Score: 62 **Number of documents:** 20