

The Recap System for Identifying Information Flow

Donald Metzler, Yaniv Bernstein, W. Bruce Croft, Alistair Moffat, and Justin Zobel
U. Massachusetts, RMIT U. , U. Massachusetts, U. Melbourne, RMIT U.

The Kremlin painted a gloomy picture of the Soviet economy, with a national debt of nearly \$500 billion -- one that is also growing faster than that of the United States. The assessment came in a speech by State Planning Committee chairman Yuri D. Maslyukov before the Supreme Soviet, or legislature, which is to consider the 1990 budget during a session that begins Sept. 26. Inflationary pressures are growing, and the deficit for 1989 is now projected at \$192 billion, up 20% from estimates given six months ago. That is four times the U.S. budget deficit in terms of the economy's ability to adjust to it, Soviet officials said. The Soviet deficit amounts to 13.8% of all the country's goods and services produced in a year.

The Soviet Union on Saturday portrayed a gloomy picture of an economy sliding toward disaster with a national debt of nearly \$500 billion and growing faster than that of the United States.

The revelation was made in a speech by State Planning Committee Chairman Yuri Maslyukov to the Supreme Soviet, or legislature.

Despite the dismal prognosis, Maslyukov's speech was markedly different from previous statements by Soviet economic authorities in that he acknowledged some key problems and outlined ways to solve them. The proposals likely will be taken up by the Supreme Soviet when it considers the 1990 budget in a session that begins Sept. 25.

Human rights activist Andrei Sakharov, speaking in the United States last week, reported the latest sad joke: The difference between an optimist and a pessimist, he said, depends on whether a person expects economic disaster to strike the Soviet Union in one, two or five years.

Soviet stores are expected to become even more barren because wages now are rising twice as fast as the production of consumer goods, according to Maslyukov's figures. Imports alone are not the answer, he said. "No foreign uncle is going to solve this problem for us," he warned.

Inflationary pressures are growing, and the deficit for 1989 is now projected at \$192 billion, up 20 percent from estimates given six months ago.

That is four times more than the U.S. budget deficit in terms of the economy's ability to adjust to it. The Soviet deficit amounts to 13.8 percent of all the country's goods and services produced in a year.

Timeline: 1988, 1989, 1990, 1991

Categories. H.3.3 Information Search and Retrieval.

Keywords. Text reuse, information flow, statistical translation.

Overview. Many kinds of documents involve reuse of previously published material. For example, typical newspaper articles combine new information with recapitulation of pertinent background.

Given a text snippet, passage, or even entire document, it is useful to be able to determine the source of the information in the passage. More specifically, the task is to seek, on a sentence-by-sentence basis, other instances of the same information. Since the emphasis is on factual information, this task is related to some aspects of question answering, but where the "queries" are the original sentences and the "answers" are transformations of those sentences. Our trust in a piece of factual information is influenced by the diversity of sources where we discover the fact.

If, however, all references to some information are derived from the same source, then our trust in the information is diminished.

Source identification can also be seen as a form of plagiarism detection. However, it is repetition of elements such as sentences (not whole documents) that is of interest in this task. In information collections such as newswires, concise statements of fact are often repeated verbatim. For example, consider the description of the eruption of Mt St. Helens as "the May 18, 1980 eruption that leveled 230 square miles, left 57 people dead or missing and sent up an ash cloud that circled the globe". While the individual elements of this statement – the date, the area affected, the number of deaths, the fact of the ash cloud – could come from a wide range of sources, when the whole is repeated it is likely that one is the source of the other, or that some third article has been used as a source by both.

The RECAP system. We have developed sentence- and document-level similarity measures for the task of finding similar information, with the intention of embedding several novel properties. In particular, we seek to:

- Favor documents with matching sentences; documents without a sentence match are low-ranked.
- Favor sentences that use the same form of words as the information provided; other sentences that are similar (in the usual IR sense) are also found, but not ranked as highly.
- Favor information-rich sentences.
- Favor documents in which there is a match for several of the sentences in the provided information.

A RECAP user enters a block of text and is shown summaries of matches, strength-of-match indicators, and a timeline showing when the matches occurred, as shown above. The latter is possible because most newswire data is annotated by date of publication. When a match is selected, matching sentences are highlighted, allowing rapid browsing of potential sources. The timeline is of particular value for the intended task; for example, it often reveals that information has been re-used by the newswire over several years.

Our prototype RECAP system provides a persuasive demonstration that source-finding is useful. The search model is, for this task, considerably more informative than the list-of-answers that is standard for search engines, and as our demonstration shows, is simple and intuitive. With a range of potential specialist applications, our software introduces a valid new form of search.

Funding. This work was supported in part by the CIIR, in part by ARDA, in part by NSF grant #CCF-0205575, and in part by the Australian Research Council.

Reference. D. Metzler, Y. Bernstein, W. B. Croft, A. Moffat, and J. Zobel. *Similarity Measures for Tracking Information Flow*. May 2005. Submitted for publication.