Derivation of Knowledge from Digital Content

A. Shelly Spearing, Jorge H. Román, HPC-1; M. Linn Collins, STBPO-RL

As electronic content proliferates, it becomes nearly impossible to fully consume and assess all of the available information. Over the past 7 years, LANL's Digital Knowledge Factory (DKF) has created a suite of digital-content-analysis tools. This combination of commercial off the shelf (COTS), Open Source, and homegrown code is used to gather, reduce, annotate, organize, synthesize, and visualize digital content for human consumption.

The tools can be applied to collections of text-based documents from virtually any source. The algorithms go beyond traditional natural language processing and statistical analysis—word-location algorithms automatically extract the gist of the content, while others annotate targeted concepts, organize documents, and calculate goodness-of-fit with respect to a specified conceptual area. Additional modules extract features, such as dates and locations, and group documents for comparative analyses. In some cases we are also able to compute the trustworthiness and/or mood of the author. When looking at larger collections, we categorize subject-matter expertise, emerging and fading trends, and distill entire collections into a variety of single-page graphical representations.

Structured information (i.e., metadata) can augment the digital knowledge to facilitate analyses of time trends, geographical colocation, and authorship, among others. Through information reduction, annotation, fusion, and organization, the analyst is able to assimilate content and form hypotheses more quickly.

A goal of the DKF project is to expedite knowledge assimilation by synthesizing digital content into a set of knowledge visualization schemes. In particular, it is hoped that an information consumer can look first at a few graphical representations of the concepts contained within thousands of pages of text, draw conclusions about the documents in aggregate, formulate hypotheses, and then focus attention on the particular documents that are relevant to the conclusions and hypotheses. Through a reduction process we focus the analyst on the important concepts and the relationships among them.

A first step in trend identification is depicted in Fig. 1, a timeline summary of 475 documents containing "Iraq" that were extracted from the White House Press Archive website. DKF tools were used to identify the top-level concepts

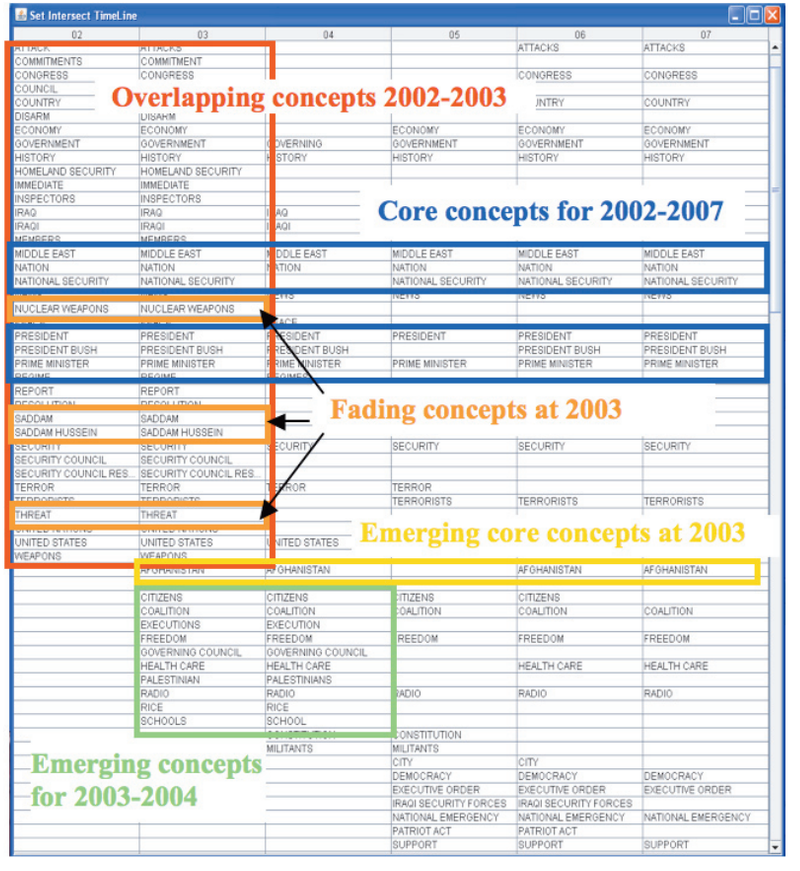

Fig. 1. Timeline summary (right) of 475 documents.

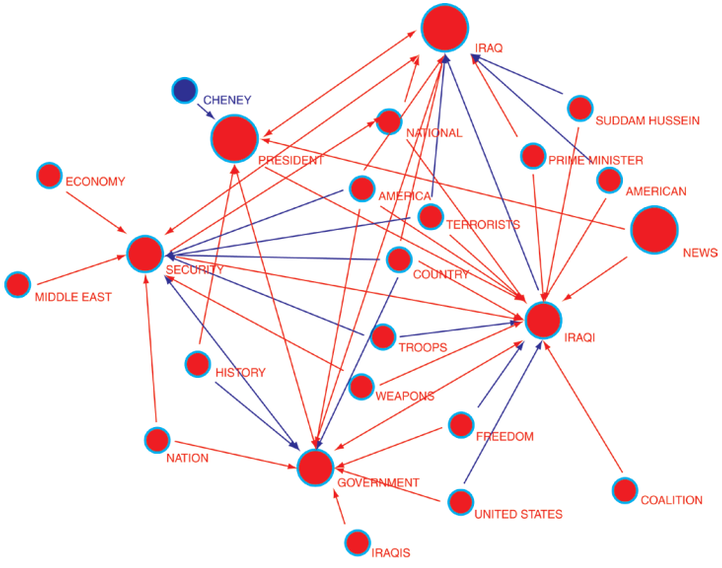

Fig. 2. Knowledge network (left) of the 475 documents in Fig. 1.

of each document, aggregate those concepts by year, and compare each year's important topics with those of other years. Overlapping concepts are those that appear in two or more consecutive years, although when they are found in all or most of the time intervals, they are considered "core" to the collection. Fading concepts are present for a few years at the beginning but not at the end of a time sequence, while emerging concepts do the opposite. In this particular example, the reader may observe that "Nuclear Weapons," "Saddam Hussein," and "Threat" were frequently found as key themes in the 2002-2003 timeframe, but not afterward. In their place, we see the emergence of "Afghanistan."

Figure 2 depicts a knowledge network representation of the same 475 speeches. In this case, secondary, or supporting, concepts are considered in relationship to the concepts they support. Top-level concepts and relationships are shown in red, second-level concepts and relationships are shown in blue. Node size and edge thickness represent frequency within this set, and direction of edge (arrows) signifies a concept relationship. Tertiary concepts and low-frequency concepts and relationships are omitted here, but can be explored using interactive tools. In the network depicted above, "President," "News," and "Iraq" are revealed as the most important concepts. The concept "nation" is supported by the concepts "government" and "security," meaning that these concepts are often found in close proximity (i.e., within the same sentence or paragraph) in speeches. "Security" is a supporting concept for many others, as depicted by the numerous edges flowing into "Security."

Through the use of these two representations, someone totally unfamiliar with world politics might quickly conclude that the source of the data is newsfeed somehow involving the President, Iraq, and national security and that, over time, the news shifted away from weapons and to Afghanistan. In other, potentially less politically prominent, contexts, the DKF tools can be used to quickly highlight changes that might warrant further human interrogation. Beyond the collection-level analysis techniques illustrated here, some other capabilities of the DKF tool suite include the targeting of specific concepts of interest and the generation of "drill down" versions of individual documents in which key and/or targeted concepts are hyperlinked and colorcoded.

The process is automated and interactive, allowing analysts to quickly focus on topics of interest in a particular set or subset of documents. DKF tools provide Knowledge Discovery capabilities out of the box, and these capabilities can be enhanced through the use of ontologies and other context-aware knowledge. Currently, we have a set of working prototypes and expect to field some of these tools to other government agencies. Our current work has focused on developing working instances of digital knowledge management (DKM) operations of the first-, second-, and third-order. We envision many more DKM operations, as these are only the core operations.

For further information contact Shelly Spearing at shellys@lanl.gov.

Funding Acknowledgments
- LANL Research Library, Emergency Situation Overview and Synthesis
- Various other programs