Current Approaches to Automated Information Evaluation and their Applicability to Priority Intelligence Requirement Answering

Brian Ulicny,
Christopher J. Matheus
VIStology, Inc.
Framingham, MA USA.
{bulicny,cmatheus}@vistology.com

Gerald M. Powell
U.S. Army Research Laboratory
Fort Monmouth, NJ, U.S.A
gerald.m.powell@us.army.mil

Mieczyslaw M. Kokar
Northeastern University
Boston, MA USA
mkokar@ece.neu.edu

Abstract **Doctrinally, Priority Intelligence Requirements (PIRs) represent information that the commander needs to know in order to make a decision or achieve a desired effect. Networked warfare provides the intelligence officer with access to multitudes of sensor outputs and reports, often from unfamiliar sources. Counterinsurgency requires evaluating information across all PMESII-PT categories: Political, Military, Economic, Social, Infrastructure Information, Physical Environment and Time. How should analysts evaluate this information? NATO's STANAG (Standard Agreement) 2022 requires that every piece of information in intelligence reports used to answer PIRs should be evaluated along two independent dimensions: the reliability of its source and the credibility of the information. Recent developments in information retrieval technologies, including social search technologies, incorporate metrics of information evaluation, reliability and credibility, such as Google's PageRank. In this paper, we survey various current approaches to automatic information evaluation and explore their applicability to the information evaluation and PIR answering tasks.**

Keywords: PIR answering, information evaluation, information retrieval, reliability, credibility

1 Introduction

Doctrinally, Priority Intelligence Requirements (PIRs) represent information that the commander needs to know in order to make a decision or achieve a desired effect. PIRs drive the military intelligence collection process and are "those intelligence requirements for which a commander has an anticipated and stated priority in his task of planning and decision making" (FM 2-0 "Intelligence", section 1-32). PIRs are a subset of the whole spectrum of information requirements, broadly speaking, that a military intelligence officer (e.g., S2) and his staff are tasked with answering.

Priority Intelligence Requirements are associated with one or more Indicators. Indicators are empirically observable variables about which information can be collected (or inferred) that would provide a (total or partial) answer to the overall PIR either directly or through analysis and inference. Each Indicator is then associated with one or more Specific Information Requirements (SIRs) that detail what information is to be collected about an Indicator. Finally, these SIRs are associated with concrete collection tasks assigned to particular personnel or sensors; assets are scheduled, units are deployed, and attempts to collect the information are made. Based on the collected information, the intelligence organization (e.g., S2 shop) produces answers to the assigned PIRs and regularly briefs the commander.

Networked warfare provides the intelligence officer with access to a vast body of information contained in multitudes of intelligence reports from intelligence assets (humans and sensors) as well as information produced by non-intelligence personnel. PIRs may involve information spanning all of the PMESII-PT (Political, Military, Economic, Social, Infrastructure, Information, Physical Environment and Time) categories. For example, the chief of coalition intelligence in Afghanistan has recently encouraged battalion S2 shops to evaluate:

> census data and patrol debriefs; minutes from shuras with local farmers and tribal leaders; after-action reports from civil affairs officers and Provincial Reconstruction Teams (PRTs); polling data and atmospherics reports from psychological operations and female engagement teams; … translated summaries of radio broadcasts that influence local farmers, …[and] the field observations of Afghan soldiers, United Nations officials, and non-gov-ernmental organizations (NGOs). This vast and underappreciated body of information, almost all of which is unclassified, … provide[s] elements of … strategic importance – a map for leveraging popular support and marginalizing the insurgency" [6].

To a significant extent, the PIR answering task is a question-answering task: queries are issued (PIRs, SIRs, and other information requirements), and information is collated and presented as an answer to the commander in a briefing. To answer a PIR, an S2 must either identify relevant information that has already been collected or task the collection of new information for an SIR. Then

the S2 must have relevant information collected, locate relevant information that has already been produced, evaluate it, analyze it, interpret it, and produce from it an answer that can be briefed and justified to the commander.

Information retrieval, broadly construed, and knowledge management thus form important elements of the PIR answering task. Behind the scenes, information retrieval technologies automatically evaluate information sources. Recent developments in commercial search technology have accelerated and become ubiquitous in civilian life. To what extent can commercial search technologies assist with the task of PIR answering? Specifically, to what extent do commercial question-answering technologies implement useful metrics of information evaluation that translate to military requirements?

1.1 Information Evaluation

NATO STANAG (Standard Agreement) 2022 "Intelligence Reports" [15] states that where possible, "an evaluation of each separate item of information included in an intelligence report, and not merely the report as a whole" should be made. It presents an alpha-numeric rating of "confidence" in a piece of information which combines a measurement of the reliability of the source of the information and a numeric measurement of the credibility of a piece of information "when examined in the light of existing knowledge".[1]

Reliability of the source is designated by a letter A to F signifying various degrees of confidence as follows:
A: Completely reliable. It refers to a tried and trusted source which can be depended upon with confidence.
B: Usually reliable. It refers to a source which has been successfully used in the past but for which there is still some element of doubt in particular cases.
C: Fairly reliable. It refers to a source which has occasionally been used in the past and upon which some degree of confidence can be based.
D: Not usually reliable. It refers to a source which has been used in the past but has proved more often than not unreliable.
E: Unreliable. It refers to a source which has been used in the past and has proved unworthy of any confidence.
F: Reliability cannot be judged. It refers to a source which has not been used in the past

Credibility: The credibility of a piece of information is rated numerically from 1 to 6 as follows:
1: If it can be stated with certainty that the reported information originates from another source than the already existing information on the same subject, then it is classified as "confirmed by other sources".[2]
2: If the independence of the source of any item of information cannot be guaranteed, but if, from the quantity and quality of previous reports, its likelihood is nevertheless regarded as sufficiently established, then the information should be classified as ``probably true''.
3: If, despite there being insufficient confirmation to establish any higher degree of likelihood, a freshly reported item of information does not conflict with the previously reported behaviour pattern of the target, the item may be classified as ``possibly true''.
4: An item of information which tends to conflict with the previously reported or established behaviour pattern of an intelligence target should be classified as ``doubtful'' and given a rating of 4.
5: An item of information that positively contradicts previously reported information or conflicts with the established behaviour pattern of an intelligence target in a marked degree should be classified as ``improbable'' and given a rating of 5.
6: An item of information the truth of which cannot be judged.

As such, the credibility metric involves notions of source independence, (in)consistency with past reports, and the quality and quantity of previous reports.

These rubrics suggest an epistemic calculus for fusing information reports by a formal reasoning system, where the evaluations are epistemic logical operators over the statements with which they are associated.

Example 1: If a source's statement *s* is classified A2 (meaning *s* has reliability A and credibility 2) and another source's statement *s* is classified as A2 as well, then both statements can be upgraded to A1 status (independently confirmed).

Example 2: From A2*p* (meaning source S states that *p* is true, with reliability A and credibility 2) and E5$\neg p$ (it is not the case that *p*, source T), we should not infer a contradiction ($p \ \& \ \neg p$), from which anything follows. Rather we should either block the inference of the contradiction, or refuse to infer every statement as a consequence, as in a paraconsistent logic [14].

Example 3: From B2*p* (usually reliable, probably true that *p*) (source S) and, independently, from source T, B4*p* (usually reliable, doubtful that *p*), we should (perhaps) infer that B3*p* (*p* is possibly true).

The set of inference rules for these operators could be completed in several ways, which is an open issue.

Unfortunately, the STANAG 2022 rubric seems to assume many conditions that are dubious in today's environment. First, the rubrics for assigning credibility

[1] The same matrix is presented in Appendix B "Source and Information Reliability Matrix" of FM-2-22.3 "Human Intelligence Collector Operations" (2006) without citing STANAG 2022. JC3IEDM [12] includes a reporting-data-reliability-code rubric that is nearly identical, with some quantitative guidance ("not usually reliable" means less than 70% accurate over time.)

[2] JC3IEDM's reporting-data-accuracy codes are nearly identical to these except that the top three categories refer to confirmation by 3, 2 or 1 independent sources, respectively. JC3IEDM also contains an additional, unrelated reporting-data-credibility-code (reported as fact, reported as plausible, reported as uncertain, indeterminate); it is not clear how it relates to the others.

are assigned with respect to "previous reports". However, we cannot assume every evaluator will have access to all and only the same reports when they do their evaluations. In a distributed environment, an S2 in one area might have access to a different set of reports than an S2 in another area, and their superior might have to reconcile the differences in their evaluations. Huge volumes of data make this difficult to do manually; hence the need for automation not only of question answering but also information evaluation.

More importantly, as the basis for combining information with confidence measures, the STANAG 2022 rubric seems to assume a reasonably small set of sources and sensors that can be independently assessed and tracked for their reliability, with novel sources providing only a small percentage of the information at any time. That is, if every piece of information is F6 (reliability and truth cannot be judged), nothing can be inferred. In today's environment, these assumptions may be reversed: novel sources may be more of the rule than the exception; determining their reliability may be infeasible; and they may provide a substantial amount of information relevant to a question.

2 Contemporary Question-Answering Technologies

The idea of question-answering by computers has a long history in Artificial Intelligence and Information Retrieval [19]. In this section, we classify contemporary information retrieval and question-answering systems by means of the way in which they internally represent information sources in order to produce an answer to a specified query. These systems go out and acquire data (usually, by following hypertext links), index its content and evaluate its quality, and then provide responses to queries about the data they contain. Our classification is based on published descriptions of the systems. In each case, we identify the representational scheme; provide examples of some familiar technologies using that representation; and provide examples of more recent, advanced applications that answer queries or retrieve information stored in that representation (Table 1). In the following sections, we will describe how these systems deal with issues of reliability and credibility, as the STANAG 2022 describes them.

Information Source Representation	**Common Application**	**Advanced Application**
Tables (Relational databases, spreadsheets, etc.)	Structured Query Language (SQL)	Wolfram Alpha
Text	Web search engines (Google, Ask, etc.)	TREC QA track; Aquaint (Advanced Question-Answering for Intelligence) systems
Tagged Text	Google Patent Search	Metacarta; Semantic MediaWiki; Palantir
Logic Statements	Prolog	Powerset (Microsoft Bing); Cyc
Trusted Teammates (Personal Knowledge)	Personal communication	Yahoo! Answers; Army ICON Shoutbox; PlatoonLeader Vark

Table 1 Question-Answering Technology by Information Source Representation

2.1 Structured Data

In computerized information systems, the use of relational databases has a long history. The vast majority of systems that store and retrieve data are based on representing the data in structured formats, in which the structure of the tables, and the significance of each column, is specified in advance. Structured Query Language (SQL) commands and queries are then used to insert and retrieve data elements in tabular form. While it has become increasingly sophisticated over the years, SQL was initially envisioned as a natural language interface to databases. In web-enabled database applications, the SQL queries and commands are mostly hidden from the user and are constructed and executed when a user fills out and submits a form on a web page.

Wolfram Alpha represents a more sophisticated version of structured data querying. Wolfram Research is the producer of the major symbolic mathematical computation engine Mathematica. The Wolfram Alpha engine sits on top of quantitative data and other reference works that have been "curated" from authoritative sources[16]. When a user queries Wolfram Alpha (Figure 1) the engine attempts to interpret the query's intent so as to produce an output format that is the most likely to satisfy that query intention (here providing both a geospatial overlay and timeline as output, automatically defaulting to data within the last 30 years), without requiring the user to formulate the underlying Mathematica query.

2.2 Text

Text search engines are the most ubiquitous technology in Table 1. Users increasingly use text search engines every day in their work and personal lives. Search engines are not strictly speaking question-answering engines, because

what they return is a ranked set of documents determined by the search engine to be relevant to the user query, not specific answers. The documents are ranked based on the frequency and position of the query terms in the document, as well as evaluations of the document's quality

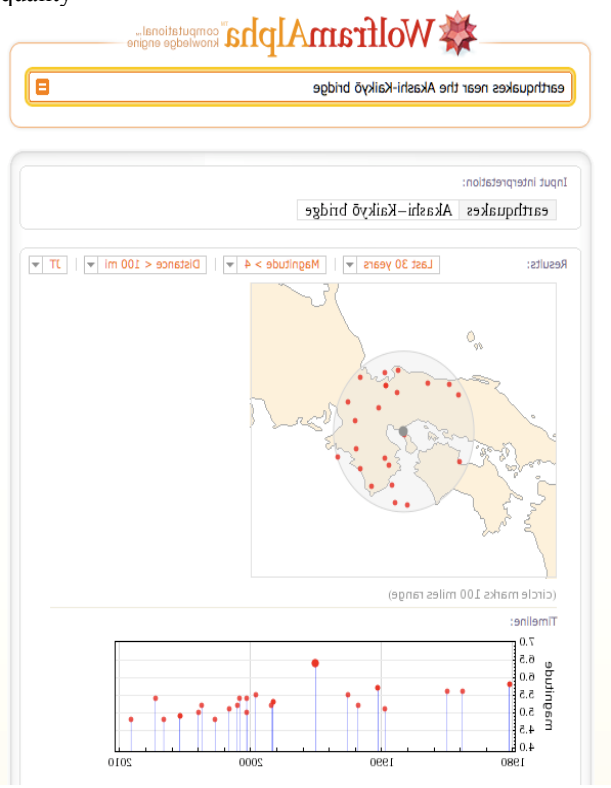

Figure 1 Wolfram Alpha output for query "Earthquakes near the Akashi-Kaikyō bridge"

as determined by Google's PageRank algorithm [13], or other measures. Sites like CNN.com and Wikipedia.org have very high PageRank (9, on a logarithmic scale of 10); an obscure blog post has very low PageRank.

Often, the user can determine the answer to a question posed as a search engine query simply by surveying the document snippets that the search engine returns, without clicking through to the documents themselves. Thus, from the snippets returned by Google one can gather that the consensus answer to the question "Where was Elvis born?" is Tupelo, MS. 'Elvis' is automatically interpreted as 'Elvis Presley' according to the highest-ranked results; no other Elvis is mentioned.

2.3 Tagged Text

By "tagged text" systems, we mean question-answering systems that operate upon semi-structured data sources: textual data to which some information about objects, properties or relationships has been identified and marked up within the document, either as metadata separate from the body of the text, or marked up inline within the text, as in this bit of HTML, which indicates that the contents of this HTML element are of the kind "date-header".

```
<h2 class='date-header'>Thursday,
      January 21, 2010</h2>
```

While the identification and markup of such data elements can occur when a document is authored, many of the information retrieval systems using tagged data as their source data representation include automated text processing in which specific types of information are identified within a text using natural language processing techniques. For example, a system might identify persons or organizations which are then inserted as markup into the system's representation of the document.

MetaCarta's technology [11] processes documents in order to identify any expression related to a location (e.g. location name or postal code), and marks up its representation of the document with geo-coordinates corresponding to that location expression. The system can then be queried for documents that contain some combination of keywords that have some geocoordinates within a specified bounding box or radius.

2.4 Logical Statements

Logic-based systems, such as Powerset, recently acquired by Microsoft and incorporated into its Bing search engine, parse all of a text into a logical representation, using sophisticated natural language processing [18]. After analyzing free text and converting it into a logic-based representation, questions can be formulated as queries over these logical clauses and returned as answers. Many formal reasoning systems today have converged on a Subject, Predicate, Object representation of logical clauses (e.g. "John kissed Mary" has subject "John", predicate "kiss", and Object "Mary", ignoring tense) as being computationally more tractable than more flexible representations.[3] Notice that in the example query below (Figure 2) clauses corresponding to various "Elvises" are highlighted, including Elvis Presley, skater Elvis Stojko, and Norwegian Elvis impersonator Kjell Elvis. Because Powerset is processing logical statements derived from a single source of data (Wikipedia), the answers do not converge on a single Elvis (Presley), as they do on the Web, where hyperlinking and network typology are used by Google to rank documents for return as well.

[3] Logical statements in triples or non-triples formats are equivalent, and they can be automatically transformed from one format to another, through a process known as 'currying' or 'Schoenfinkalization'.

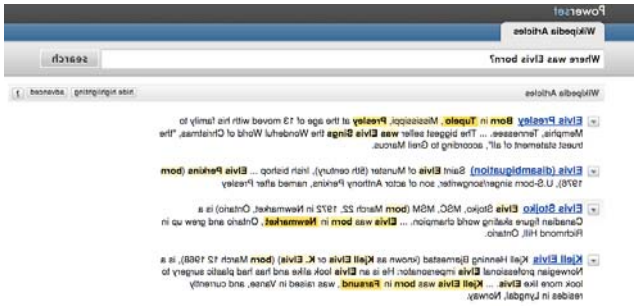

Figure 2 Powerset output for "Where was Elvis born"

2.5 Trusted Teammates

Finally, the last representation of answer sources is knowledge contained within the heads of what we call "Trusted Teammates". Surely, the oldest technique for question- answering is simply to ask someone you trust who you believe knows the answer.

In the context of the Web, this has been updated in services such as Yahoo! Answers (answers.yahoo.com) to allow users to pose questions to a community of online respondents, who provide answers asynchronously. The users can then use statistics compiled on the various respondents in order to assist in evaluating both the source and the content of the answer provided (the number of answers they have provided, their areas of expertise, the amount of positive feedback they have received, etc. Figure 3). A simpler question-broadcast service is incorporated as the "Shout Box" function in the Army's Intelligence Center Online Network (ICON) [3].

Vark (Vark.com), recently acquired by Google, is another social question-answering application. In Vark, users do not seek out questions to answer in a central repository; Vark attempts to automatically identify the person in a user's social network (gleaned from their Facebook, Twitter, IM (instant messenger) contacts and the like) that is most likely to be able to answer the question. This user-respondent quality metric is computed as the weighted cosine similarity over a feature vector that

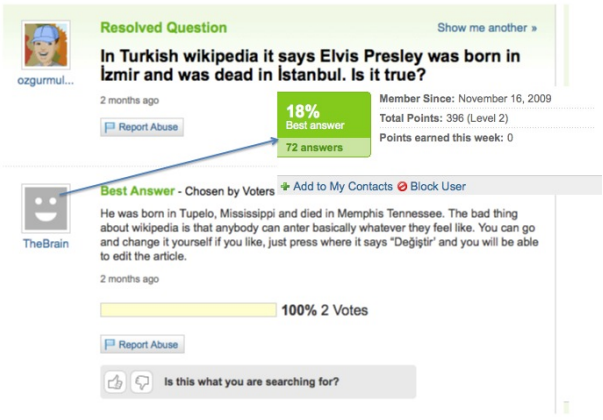

Figure 3. Yahoo! Answers

includes both social network proximity and overlap metrics as well as metrics of topic overlap (vocabulary and stated interests) and demographic overlap [8]. The service manages connecting the asker and respondent and handling their interaction.

SRI's iLink framework [4] is a similar social search system that returns user queries and experts in response to a query. It also suggests ongoing discussions to a user on the basis of his or her past participation and interests. Finally, it has an automated moderation function that prompts a user to participate in a discussion based on their interests and past participation. This functionality has been deployed in the U.S. Army's PlatoonLeader website (platoonleader.army.mil), a knowledge-sharing site for past, current and future platoon leaders in the US Army, organized around topics such as leadership and fitness.

3 Current Approaches to Reliability

The STANAG 2022 standard for evaluating reliability is based on past accuracy: a source is considered reliable to the extent that its past statements have been true. Trust is a correlate of reliability: it is rational for someone to trust a source or system to the extent that it is reliable. (In human behavior, trust undoubtedly has many irrational components as well.)

It is not clear how source reliability is tracked and monitored by human S2s in practice today. It is possible that a running score of statements to verified true statements is maintained for each source and sensor. Comparing past reliability across sources poses some problems, however. The truth-conditions of statements differ considerably in scale: it is one thing for a source to correctly state that "Mr. X is in the café"; it is quite a different matter for a different source to correctly state that "The army is preparing to invade Country Y". On the basis of only these true assertions, the two sources would be considered equally reliable, although the importance of their statements is quite different in scale.

In a networked environment like the contemporary operating environment, an analyst is exposed to many novel sources of information across PMESII-PT categories and has very little ability to check their reliability directly. The STANAG 2022 standard requires that novel information sources be given an unknown reliability rating, but that seems unreasonable. It treats all novel information sources as equally suspicious, when in fact most users are comfortable with indirect estimates of unknown data reliability.

Structured data services such as Wolfram Alpha deal with the reliability issue by only "curating" data from highly reliable sources, such as standard reference works or official, government data. Curated depositories like this exist in military intelligence contexts as well. Sources that are included in such a repository, even if they are novel to the evaluator, can be considered highly

reliable. In such cases, if one trusts the curator, one doesn't need to independently establish the reliability of a source; the fact that it is endorsed by the curator is enough.

In standard contemporary text-based information retrieval models, an information quality metric is computed for all documents in addition to the relevance metric, matching a document to the specific information need expressed by the query. This is done independently of assessing their reliability directly. That is, contemporary search engines consider two factors when they return a document in response to a query: a representation of what the document is about, usually based on the frequency distribution of terms in a document and across other documents; and a representation of how good the document is, based on an analysis of network properties. Google, that is, does not fact-check the content of a site to evaluate its information; it uses network properties that it believes are highly correlated with information quality or reliability as a correlate of reliability; these rankings can change as user hyperlinking behavior changes.

Google's PageRank algorithm [13] and variants to it have been highly successful in presenting users with reliable information without direct fact-checking. The PageRank algorithm calculates a document's quality recursively, weighing inlinks from high-quality documents (those that are themselves pointed to by high quality documents) more highly. The PageRank algorithm can be understood as computing the likelihood that a random web surfer will end up on a particular document given that, for each document, the web surfer tends to jump to a new document some percentage of the time (standardly, a 15% likelihood to jump is employed as the so-called damping factor, defining the propensity to continue to a new page). This algorithm is recursive and typically computed for only a small number of iterations, because it would be too computationally expensive to extend the computation to the entire Web graph. Hyperlinks are assumed to be made by disinterested parties, not for the sake of PageRank itself. "Link-farming" to inflate PageRank is ferreted out.

Many other highly successful information evaluation technologies have evolved that all rely, to one degree or another, on network analysis properties: centrality, overlap, distance and so on. For intelligence reports themselves, metrics like Google PageRank are less applicable, because the intelligence reports are not hyperlinked

However, networked-based metrics like PageRank are clearly applicable to many open-source and unclassified data sources, such as news sites and blogs, to provide an estimate of reliability, even when they have not been encountered previously.

Blogs, for example, are an important venue for political mobilization and recruitment. Attributions of responsibility for a terrorist bombing may appear in terrorist blogs. Technorati, a blog search engine, uses the relatively simple metric of in-link centrality, the number of links from other blogs over the last six months, as their blog quality metric. Some of the present authors have shown that a metric combining both Technorati authority and reader engagement, as measured by blog comment counts, as well as accountability-enhancing profile features, outperforms both PageRank and Technorati Authority alone in ranking social-political blogs by their credibility [17], which correlates with their influence. Similar network-based metrics for estimating reliability apply to source documents for tagged text and logic-based solutions as well.

Social search engines such as Yahoo!'s question answering service track user feedback on respondents as a metric of reliability. Yahoo! tracks the number of users who have rated a respondent's answer as the best provided (which is different from directly confirmed). We have seen that the social search service Vark computes a respondent quality metric corresponding to the likelihood that a respondent will provide a satisfactory response to the user based on social network, demographic similarity, profile similarity, response metrics (speed, length), and so on [8].

Social search metrics such as those incorporated by Vark are surely applicable to estimating reliability among teammates or coalition partner information sources, such as non-governmental organizations (NGOs) and the like, whose information is likely to be important in full spectrum counterinsurgency environments. Such metrics are also applicable to estimating the reliability of unfriendly or potentially hostile sources with respect to their social networks. A highly central figure has more authority, and is probably more likely to be reliable than a marginal figure in a social network, at least with respect to information that involves that network or its participants. Inference can be made in the other direction, as well: a highly accurate source may be inferred to be central to a social or knowledge network, contrary to appearances.[4]

We conclude, then, that the network-theoretic metrics used in civilian information retrieval applications, should be investigated for systematically estimating source reliability in military intelligence contexts as well, if tracking source reliability directly is impractical or unfeasible. These networks apply to the hypertext graph of the Web, the graph of the blogosphere, the Twitter graph, social networks (Vark) and so on. They are especially useful in providing an estimate of the reliability of previously unknown sources.

[4] Witness the case of Helen Duncan and the HMS Barnham in WWII Britain. Duncan was suspected of conspiracy for her knowledge of British wartime secrets, which she said were communicated to her supernaturally in séances, but authorities suspected she was a spy. http://www.webatomics.com/jason/-barhamconspiracy.html

One difficulty here is making network-based reliability estimates commensurable across networks of different types. For example, the reliability of a blog may be estimated with respect to the blog network; and the reliability of a Twitter user may be estimated with respect to the Twitter network, but it is not clear that a blogger who ranks in the nth percentile according to a metric for blogs is as reliable as a user who ranks in the nth percentile according to a metric for Twitter users or for Vark users. Scaling metrics to network size may be important to make the metrics commensurable.

4 Current Approaches to Credibility

The contemporary operating environment poses many challenges for the STANAG 2022 rubric on credibility. STANAG 2022 credibility guidelines determine a piece of information's credibility on the basis of (i) sameness of information, (ii) confirmation by an independent source, and (iii) consistency with previous reports.

STANAG 2022's highest credibility ranking goes to information that is independently confirmed. Suppose an analyst sees two Twitter status updates, from two different accounts A and B, each saying "The Archduke has been shot". It seems premature to say that the two Twitter updates are ipso facto independent. Both Twitter updates might merely be rebroadcasting what a mutual contact, C, had said previously. On social media platforms, it is often possible to trace how information flows from one user to another directly by means of hypertext trails, "retweet" or "hat tip" citations, timestamps and other mechanisms

In a network of sources, independent confirmation must require independence of sources, not primarily independence of content. No two sources are independent in the sense that no path exists from one source to another through the social graph, and any piece of information is likely to propagate from one node to another over time. Rather, independent confirmation must mean that if A and B both report the same thing, and A and B do not have a shortest path between them closer than the average shortest path length between any two nodes in the social network and there is no source C in the network who reports the same thing that has a shorter path between both A and B than A and B have to one another, then A and B are independent confirmations of one another.

In the context of social media platforms such as Twitter and Ushahidi[5], on which sources can proliferate freely and contribute information anonymously, it becomes important to identify how many users are saying the same thing. Metrics for this range from tracking common URLs to computing n-gram overlaps between texts (Rouge) to tracking parts of the same quotation through news stories [9] to automatically binning messages by content or sentiment through statistical analysis. In [2], the authors present a simple metric of report sameness using ontologies, but they do not provide for the automatic detection of inconsistencies. In [7], the authors provide a sophisticated method for estimating the proportion of texts of the same type in a corpus (e.g. Twitter updates expressing the same attitude about the State of the Union) without training individual classifiers for each type.

Finally, the STANAG 2022 credibility rubric depends on consistency with prior reports. An item p is consistent with previously gathered information I if it is not possible to infer a contradiction from I and p jointly. For example; the statements

(1) A was born in the same town as B;

(2) B was born in Latvia

(3) A is a native of the UK

are not consistent since one can infer a contradiction from their union: A both was and wasn't born in the UK. No single premise directly contradicts another.

It is a mistake to overvalue temporal information priority. There is a human tendency to disregard or diminish the significance of information that doesn't fit (or even blatantly contradicts) our beliefs, hypotheses, or mental models of situations. STANAG 2022 should not be taken to prioritize coherence with the earliest reports; rather, it says that the largest set of internally consistent reports on a subject is more likely to be true, without independent evidence. It is a military truism that "the first report is always wrong",[6] so a bias towards coherence with the first report on a subject should be rejected.

In structured data question-answering systems, data coherence is enforced through *integrity constraints* on data input. Such integrity constraints require that, for example, every individual must have a social security number, no two distinct individuals can have the same social security number, and no single individual can have more than one birthdate. In this way, the system prevents new, inconsistent information from being input. Such constraints may be too strict for military intelligence applications, however, because they require deciding between two possibly correct pieces of information at data entry time, and this may not be known until later.

Some logic-based systems provide for the automatic inference of inconsistencies based on a set of facts encoded as RDF (Resource Description Framework) or OWL (Web Ontology Language) triples and an ontology expressing constraints on how specific individuals and classes are related, using a logic with a tractable set of inference rules [10]. Logic-based systems do not have integrity constraints in the same way that structured data systems do; it is possible to say that every individual must

[5] The Ushahidi platform (ushahidi.org) combines a map overlay with the ability to post reports by location, via cell phone texts or from Twitter or anonymously from the web. It has been used to monitor election fraud in Afghanistan and response to the 2010 Haitian earthquake.

[6] LTG (Ret) Ricardo S. Sanchez, Military Reporters and Editors Luncheon Address. 12 Oct 2007. http://www.militaryreporters.org/sanchez_101207.html

have exactly one social security number, for example, but the system need not be able to provide it, at least for those logic-based systems that incorporate an open-world, rather than closed-world assumption. With a logic-based system, it would be possible to infer the inconsistency between a set of reports about where A was born from (1), (2), (3) given an ontology that encoded the relevant information about Latvia, the UK, and the relations "born in", "native of", and "same town". Even a trivial example like this involves a number of unstated axioms that would have to be captured: for example, that a town is a part of a country; thus, if you are born in a town, you are born in the country of which that town is a part, and so on.

Pure text-based question-answering systems have used consensus-based answers to identify answers to factual questions in a textual corpus. The AskMSR system [1] identified the most frequent phrases proximate to query terms in highly ranked documents as the answer to a query. Leveraging data redundancy in raw documents, rather than curated reports, helped the system to provide more accurate answers. Such systems are less useful if the correct answer can change with time.

While social question-answering systems incorporate metrics for reliability or source quality, we are not aware of social search systems that attempt to validate a respondent's answer by calculating its consistency with a body of prior knowledge. One exception (although not really a social search system, per se) is the winning team from MIT at DARPA's Network Challenge, in which ad hoc teams, recruited and interacting for most part via social media, competed to identify the location of ten balloons placed across the continental US. Teams were competing for money, and substantial disinformation from other teams was encountered. The MIT team evaluated the proximity of a balloon reporter's IP address to the reported location of a balloon, among other factors, in evaluating a report's credibility [13].

5 Research Gaps

The foregoing shows that, for the range of sources for which intelligence analysts in a networked environment must provide evaluations across PMESII-PT categories, direct assessments of their reliability may not be feasible. Analysts may have to rely on estimating source reliability based on metrics derived from network properties. Conversely, the profusion of data sources in a networked environment makes the establishment of independent confirmation and influence tracking more difficult. Automated techniques for identifying sameness of reports and assessing their consistency have been developed.

The open research gaps that remain include: (1) how best to map network-based reliability metrics to STANAG 2022 reliability codes; (2) how to make reliability metrics derived from networks of different scales commensurable and commensurable with non-estimated reliability metrics; (3) how to automatically reason with information that has been assigned STANAG 2022 evaluation codes; (4) how to efficiently identify independent confirmation of reports in social media; and (5) how to tractably identify inconsistent new reports; and (6) how to adjudicate inconsistencies among reports automatically.

6 Conclusion

We have described the PIR answering task and the STANAG 2022 standard for intelligence information evaluation of reliability and credibility that should inform PIR answering, doctrinally. We described various approaches to automated question-answering applications in non-military contexts and described their approaches to reliability and credibility. Our purpose has been to call attention to the work in quantitative information evaluation being done in the field of information retrieval to spark cross-fertilization with research in information fusion. Such techniques are better suited to the contemporary operating environment in which large numbers of novel intelligence sources are encountered whose reliability and credibility would be impractical to directly assess and track.

Note: This paper does not represent an endorsement by the Army Research Laboratory of any of the commercial products discussed.

References

[1] Banko, M. et al. AskMSR: Question answering using the worldwide web. *2002 AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*

[2] Besombes, J. Nimier, V., and Cholvy, L., Information evaluation in fusion using information correlation," in *Information Fusion, 2009.* July 2009.

[3] Chunn, S. The Intelligence Center Online Network: walkthrough of a KM system. Military Intelligence Professional Bulletin. January, 2008

[4] Davitz, J.; Yu, J.; Basu, S.; Gutelius, D.; Harris, *iLink:* search and routing in social networks. In *KDD '07.*

[5] Flatow, I. US Government Sponsors High-Stakes Balloon Hunt. Talk of the Nation (NPR Radio). 12/11/09.

[6] Flynn, MG M. T., Pottinger, Capt. M., USMC, Batchelor, P. D., "Fixing Intel: A Blueprint for Making Intelligence Relevant in Afghanistan", Center for a New American Security (CNAS) Working Paper. Jan 4, 2010.

[7] Hopkins, D., King, G., A Method of Automated Nonparametric Content Analysis for Social Science, *Am. J. of Political Science* 54, 1 (January 2010): 229-247

[8] Horowitz, D., Kamvar, S., Anatomy of a Large Scale *Social* Search Engine. WWW2010, Raleigh, NC. 2010.

[9] Leskovec, J.; Backstrom, L.; Kleinberg, J. Meme-tracking and the dynamics of the news cycle. In *KDD '09.*

[10] Matheus, C., *Practical OWL 2 RL Reasoning Via Fast Forward-Chaining Inference Engines*. Semantic Technology Conference, San Jose, CA, June 14-18, 2009

[11] Metacarta http://www.metacarta.com/products-platform-information-retrieval.htm (Retrieved Jan, 2010)

[12] Multilateral Interoperability Programme. THE JOINT C3 INFORMATION EXCHANGE DATA MODEL (JC3IEDM Main). Version 3.0.2. May, 2009.

[13] Page, L. Brin, S., Motwani, R., Winograd, T., (1998) The pagerank citation ranking: Bringing order to the web. Report, Stanford Digital Library Technologies Project.

[14] Priest, G., Tanaka, K., "Paraconsistent Logic", *The Stanford Encyclopedia of Philosophy (Summer 2009 Edition)*, Edward N. Zalta (ed.).

[15] STANAG 2022 (Edition 8) Annex. North Atlantic Treaty Organization (NATO)

[16] Talbot, D., "Search Me: Inside the launch of Stephen Wolfram's new "computational knowledge engine"." *Technology Review*. July/August 2009

[17] Ulicny, B., Matheus, C., Kokar, M., *Metrics for Monitoring a Social-Political Blogsophere: A Malaysian Case Study.* IEEE Internet Computing, Special Issue on Social Computing in the Blogosphere. March/April 2010.

[18] Van den Berg, M., et al, FACT-BASED INDEXING FOR NATURAL LANGUAGE SEARCH. US Patent Application. Publication number: US 2009/0063550 A1

[19] Waltz, David L. (1975) Natural-Language Question-Answering Systems. Clinic on Library Applications of Data Processing, April 27-30, 1975, ed. F.W. Lancaster. Urbana, Il: Graduate School of Library Science: 137-144.