Evaluation Metrics for Nuclear Forensics Search

Fredric C Gey
Electra Sutton
University of California, Berkeley
Inst. for Study of Societal Issues
Berkeley, CA 94720-5670
+1-510-292-8421
{gey, electra}@berkeley.edu

Charles Wang
Ray R Larson
University of California, Berkeley
Information School
Berkeley, CA 94720-4600
+1-510-642-6046
{charleswang, ray}@ischool.berkeley.edu

Chloe J Reynolds
University of California, Los Angeles
IT Services | 3327 Murphy Hall
Los Angeles, CA 90095-1434
+1-310-206-1621
creynolds@it.ucla.edu

ABSTRACT
Nuclear forensics search is an emerging sub-field of scientific search: Nuclear forensics plays an important technical role in international security. Nuclear forensic search is grounded in the science of nuclear isotope decay and the rigor of nuclear engineering. However two aspects are far from determined: Firstly, what matching formulae should be used to match between unknown (e.g. smuggled) nuclear samples and libraries of analyzed nuclear samples of known origin? Secondly, what is the appropriate evaluation measure to be applied to assess the effectiveness of search? Using a database of spent nuclear fuel samples we formulated a search experiment to try to identify the particular nuclear reactor from which an unknown sample might have came. This paper describes the experiment and also compares alternative evaluation metrics (precision at 1, 5 and 10 and mean reciprocal rank) used to judge search success

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: Information Search and Retrieval—retrieval models, search process**.**

General Terms
Experimentation, Performance, Measurement

Keywords
Nuclear Forensics, Evaluation Metrics, Scientific Search.

1. INTRODUCTION
According to [Mayer, Wallenius and Fanghänel 2007] "Since the beginning of the 1990s, when the first seizures of nuclear material were reported, the IAEA (International Atomic Energy Agency) recorded more than 800 cases of illicit trafficking of nuclear or other radioactive materials." Security agencies worldwide continue to work to prevent nuclear terrorist incidents from happening. The two aspects of prevention are detection and forensics. Millions of dollars are being spent on improvement of devices to detect contraband radioactive material which might be hidden in shipping containers. The flip side of detection is forensics – if a significant amount of smuggled nuclear material is seized, can its origin be traced to both track down the would-be terrorists and to prevent further smuggling activities [IAEA 2002, APS/AAAS 2008, GAO 2009]. To do this, a seized sample can be analyzed to ascertain its "nuclear signature" which can be compared to an archived digital library of nuclear signatures which have been abstracted by radio-chemical analysis of a large number (tens of thousands) of nuclear samples from uranium mines or nuclear reactors worldwide.

2. NUCLEAR FORENSICS SEARCH
Given a nuclear sample obtained from whatever process (interdiction, for example), the problem is to identify its source. Such identification requires clues to match against a dataset of samples for which sources and compositions have been identified. The process, abstractly, is not that different from matching fingerprints or DNA samples from a crime scene – both require a library against which the match will be made, and both require specialized matching technologies which execute the search. In the case of nuclear forensics, the library will consist of radioactive samples and their digital signatures obtained by radiochemical analyses. For the example of nuclear weapons grade material, the commonly found isotopes are highly enriched uranium (>90% ^{235}U) and plutonium (~93% ^{239}Pu).[1] The signatures of both isotopes can be characterized by their daughter isotope production from the nuclear decay process. [Gey et al 2012] describes the general search process as a temporal directed graph matching problem. In that paper and our experiments so far, temporal effects have been ignored. This is not unreasonable considering the half life of ^{235}U is 704 million years and of ^{239}Pu is 24,100 years.

3. NUCLEAR FORENSICS DATA
3.1 Spent Fuel Rod Measurements
SFCOMPO is a database of spent nuclear fuel (fuel rods from a nuclear reactor after the energy has been extracted by the nuclear fission process) measurements. The data has been carefully vetted and deemed reliable by nuclear engineering experts and has been released to the public via the Organization for Economic Cooperation Nuclear Energy Agency (OECD-NEA) web site. [2] The process by which the samples are measured (the geometry of where the sample has been drilled and extracted from the fuel rod) is described on the web site. The data consists of 274 samples from 14 nuclear reactors (some no longer in operation) in four countries (Germany, Italy, Japan and the USA). There are a variable number of samples from each reactor, ranging from two for the Genkai-1 reactor in Japan to 39 for the Trino Vercellese reactor in Italy. Each sample has a variable number of isotope, isotope ratio and burn-up measurements, ranging from one measurement for Europium 155 (^{155}Eu) to 261 measurements (a measurement is found in almost all samples) for two Uranium ratios (^{235}U/^{238}U and ^{236}U/^{238}U). The total number of measurements is 10,339.

[1] http://en.wikipedia.org/wiki/Weapons-grade#Weapons-grade_plutonium

[2] http://www.oecd-nea.org/sfcompo/

3.2 A Nuclear Forensics Search Experiment

We developed a nuclear forensics search experiment using the SFCOMPO database. Our experiment ignored all temporal effects on measurements and measurement rations. We understand that in an applied setting temporal decay would have to be taken into account, but we invoke this simplifying assumption to get at core issues in the utility of the database for search. At the same time, the crude method may provide a more realistic simulation that would seem at first consideration. Robel and Kristo [2008] noted findings from two separate experiments claiming that decay effects did not alter reactor origin determinations. "

The goal of our experiment was to determine whether a sample and its constituent measurements can be used to identify which reactor the sample came from. The structure of the search experiment is thus:

1) A single sample is removed from the set of samples in the database. This sample becomes the "query sample" and all other samples are the "known samples", or "document samples" to invoke search terminology.
2) A similarity matching algorithm is developed, using a leave-one-out approach, which matches the measurements in the query sample with each of the measurements in each document sample. This match results in a number between zero and 1 called a Retrieval Status Value, or RSV. (Ideally the RSV is an estimate of a matching probability).
3) Document samples are ranked by this similarity value.
4) Relevance of the document sample to the query sample is assessed as follows: If a document sample comes from the same reactor as the query sample, then the document sample is judged relevant. Samples not coming from the query reactor are judged irrelevant.
5) The usual information retrieval evaluation measure of precision can be calculated.

3.3 The Similarity Matching Algorithm

In order to explain the algorithm, we again describe the dataset more abstractly. The SFCOMPO dataset contains 274 nuclear material samples taken from 14 nuclear reactors. Each of the 274 samples has up to 113 measurements associated with it. That is to say, the dataset has 274 rows and 113 columns. The "measurements" are a) the quantities present in the sample of isotopes and isotopic ratios, and b) burnup values. Because conducting such measurements is expensive, many samples' measurements are missing values (the measurements were not conducted). . Figure 1 is a frequency distribution of the count of samples for the top ten measurements.

Overall, the population density for rows is low (mean = 38 out of 113 measurements present per sample, 34%). The population density for columns is slightly higher (mean = 91 out of 247 measurements per sample, 37%).

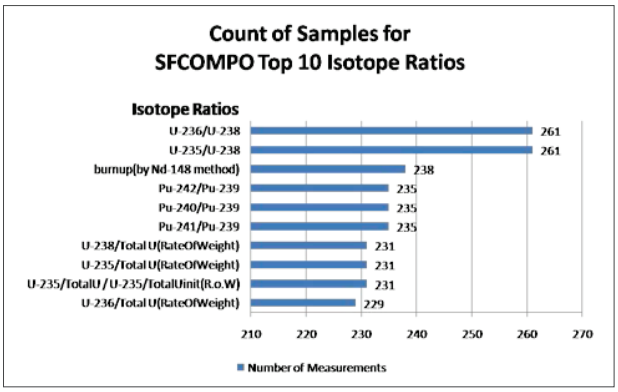

Figure 1: Top 10 Isotope Ratio Measurements in the SFCOMPO Spent Fuel dataset, by number of measurements

We extract one sample (at a time) from the database which becomes the **query sample**. The remaining 273 samples left in the database at that time are the **known (document) samples**. A sample **pair** is a pair of two samples, one of which is the query sample, the second of which is a known sample. A **column_in_common** is a column for which both samples in a pair have values, i.e. neither column is missing values.

Our naïve algorithm (in the sense that it ignores nuclear decay over time) compares the differences between columns of the query sample (input) and the columns of each known sample, ultimately creating a list of the top 10 most similar "known sample" results (output). This process works much like an internet search engine (i.e. Google). However, instead of a search term, a user would enter a query that consisted of up to 113 isotopic measurements of an interdicted nuclear material. This set of isotopic measurements of this interdicted sample is the "query sample." In lieu of getting ranked website/documents as results, the user will receive a list of relevant "known samples." The results will also display the reactor from which each result/known sample originated. In this way, we aim to detect the probable reactors of origin for interdicted nuclear samples.

First, we looked at the **range** (maximum value over the entire 274 samples minus the minimum value) for each column. Then we calculated weighted column distances for each column of each pair.

For a column x, for a pair, we compute

$$\textit{weighted column distance} = |x_q - x_k| / \textit{range of column } x$$

where x_q is the column x value of the query sample and x_k is the column x value of the known sample

Next, we compute the retrieval status value for each of the 273 pairs, which is 1 minus the root mean square of the weighted column distances for a pair, for all n columns_in_common. Thus for a known sample, the *retrieval status value (RSV)* is

$$\textit{retrieval status value, or RSV} = 1 - \sqrt{1/n \times (x_1^2 + x_2^2 + \cdots + x_n^2)}$$

where n is the number of weighted column distances for a pair and x is the weighted column distance of a column-in-common for a pair.

3.4 Results

Utilizing all samples in the database results in 274 queries, where each query sample is matched against the other 273 samples. Since a premium is placed upon correct ranking at the top of the list, we report only precision in the top ten samples retrieved. **Table 1** (after referrences) summarizes these results by reactor. Overall Precision at 10 of our naïve information system was 34%. Individually P@10 ranged from 1.0 to 0.06. However, limiting the precision to rank 10 limited our maximum possible precision by definition. These 274 samples came from 14 reactors and the number of sample per reactor ranged from 1 to 39 samples per reactor. For example, the H.B.Robinson-2 reactor had 7 samples. If one sample is used as a query then a maximum of 6 relevant document samples can appear in the top 10 rankings. Since 6 is less than 10 (the precision rank), the maximum possible precision for a query sample taken from this reactor would be 0.6 For each reactor we also compute the ratio of actual p@10 over maximum possible p@10. Finally, although comparison to random retrieval is never done in text retrieval experiments because the probability of selecting a relevant document at random is infinitesimal, for this search problem the number of samples in the entire collection is small enough to make comparison to random retrieval a reasonable task.

The last column in Table 1 shows actual precision at 10 compared to random retrieval. Averaged over all sample queries, our match performs nearly five times better than random retrieval would.

3.5 Alternative Metrics

Because of the lack of coherence of precision at 10 caused by insufficient 'relevant document' samples, we have recently also computed alternative metrics for evaluation of this nuclear forensics search process. Except for the Genkai-1 reactor, all other reactors will have at least five relevant samples in the document set for a query sample. Thus precision at 5 seems a good measure, as well as mean reciprocal rank (MRR) commonly used in evaluation of question answering systems [Voorhees 1998]. In addition, considering the nature and urgency of the international security aspects of this problem, we would also want to evaluate the 'best first' result of our search. Thus precision at rank 1 is another important metric to be computed.

Figure 2 displays all of these metrics for the 14 nuclear reactors in the SFCOMPO database. It is notable that P@ 5 and P@10 track quite well, while precision at rank one gives better results for four reactors and worse results (e.g. zero) for three reactors. Interestingly, P@1 and MRR also track well.

4. SUMMARY

This paper presents *nuclear forensics* a new application in the area of scientific search. The area has international security importance and also presents interesting challenges to develop new search and evaluation methodologies. The approach we have described, similarity matching, is not the only approach. Another interesting approach is to cast the nuclear forensics matching problem as an automatic classification problem [Robel and Kristo 2008]. The appeal to the Information Retrieval research community is to see if additional models can be founded to apply to nuclear forensics discovery.

5. ACKNOWLEDGMENTS

This research was sponsored by the USA National Science Foundation (NSF) under Grant # 1140073, entitled "ARI-MA: Recasting Nuclear Forensics Discovery as a Digital Library Search Problem." We thank Bethany Goldblum and David Weisz (UC Berkeley Nuclear Engineering) and M.C. (Mikey) Brady (Pacific Northwest National Laboratory, Richland Washington USA) for helpful discussions.

6. REFERENCES

[APS, AAAI 2008] American Physical Society (APS)/ American Association for the Advancement of Science (AAAI) Joint Working Group: **Nuclear Forensics: Role, State of the Art, Program Needs,** 2008.

[Ehmann and Vance 1991] W. D. Ehmann and D. E. Vance, **Radiochemistry and Nuclear Methods of Analysis**, John Wiley and. Sons, New York, 1991

[GAO 2009] General Accountability Office (GAO) report, **Nuclear Forensics: Comprehensive Interagency Plan Needed to Address Human Capital Issues,** *available* at http://www.gao.gov/new.items/d09527r.pdf

[Gey et al 2012] F Gey, C Reynolds, R Larson and E Sutton, "Nuclear Forensics: A Scientific Search Problem," Presented at LWA 2012, Dortmund, Germany, September 2012.

[IAEA 2002] International Atomic Energy Agency (IAEA) Staff Report, "Tracing the Source: Nuclear Forensics and Illicit Nuclear Trafficking," October 2002.

[Mayer, Wallenius and Fangänel 2007] K. Mayer, M. Wallenius, T. Fanghänel, "Nuclear forensic science — From cradle to maturity," *Journal of Alloys and Compounds 444–445 (2007)* 50–56.

[Robel and Kristo 2008] M. Robel and M. J. Kristo, "Discrimination of source reactor type by multivariate statistical analysis of uranium and plutonium isotopic concentrations in unknown irradiated nuclear fuel material," *Journal of Environmental Radioactivity 99 (2008)* 1789-1797.

[Voorhees 1999]. "TREC-8 Question Answering Track Report." in *Proceedings of the 8th Text Retrieval Conference.* pp. 77–82.

Table 1: Results for precision at rank 10 by reactor

Reactor Name	Reactor Country	Number of Measurement Sets	Actual Precision@10 (per reactor)	Max Possible Precision@10	Actual / Max Possible Precision	Random Expected Precision	Actual / Random Precision
JPDR	Japan	30	**1.00**	1	1.00	0.11	8.96
Monticello	USA	30	**0.85**	1	0.85	0.11	7.62
Tsuruga-1	Japan	10	**0.53**	0.90	0.59	0.04	14.25
Trino_Vercellese	Italy	39	**0.24**	1	0.24	0.19	1.27
Fukushima-Daini-2	Japan	18	**0.21**	1	0.21	0.07	3.14
Takahama-3	Japan	16	**0.16**	1	0.16	0.06	2.69
Fukushima-Daiichi-3	Japan	36	**0.16**	1	0.16	0.13	1.20
Obrigheim	Germany	33	**0.15**	1	0.15	0.11	1.40
Genkai-1	Japan	2	**0.10**	0.10	1.00	0.01	13.32
H.B.Robinson-2	USA	7	**0.09**	0.60	0.15	0.03	3.47
Cooper	USA	6	**0.07**	0.50	0.14	0.02	3.14
Gundremmingen	Germany	15	**0.06**	1	0.06	0.06	1.00
Mihama-3	Japan	9	**0.06**	0.80	0.08	0.03	1.76
Calvert_Cliffs-1	USA	9	**0.06**	0.80	0.08	0.03	1.79
Overall		273	**0.34**	0.84	0.37	0.07	4.86

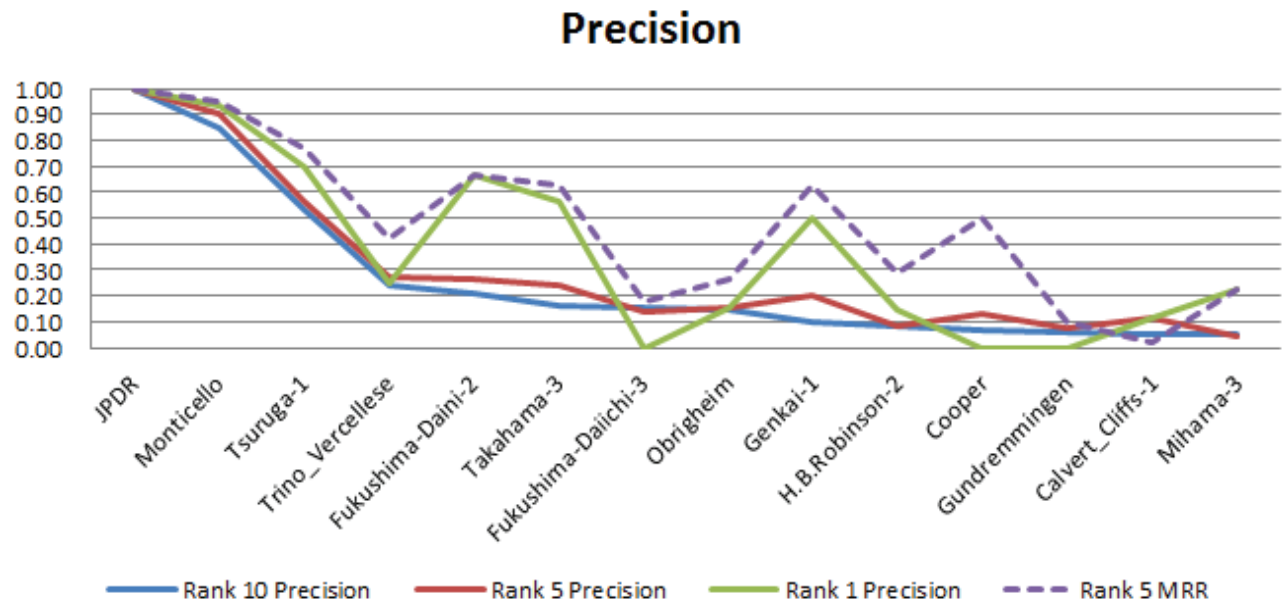

Figure 2: Precision@1,5,10 and MRR by reactor