MINING MEASURED INFORMATION FROM TEXT

Arun Maiya, Dale Visser, and Andrew Wan

Many DoD problems involve the search and exploration of scientific and technical documents. In our work, we describe an approach to extract *measured information* from text (e.g., a 1370°*C* melting point, wavelengths greater than 2.4μm). We then present MQSEARCH, a search engine that leverages these extractions to provide full support for searches based on *measured information*. Together, these capabilities better allow DoD to make sense of large and unfamiliar sets of technical documents.

MQSEARCH is a facet-based navigation system that allows users to navigate large document sets based on measured quantities, measured properties, and the topics and themes with which they are associated. To the best of our knowledge, no other search engine in existence fully supports such a capability.

Scientific and technical documents describe methods and results using measured quantities: a numeric value paired with a unit of measurement. Examples of text snippets containing such measured quantities include the following:

- *average gravity curvature* $\zeta=(1.3999\pm0.003)x10^{-5}s^{-2}m^{-1}$
- *12°C melting point*
- *distance from Earth to the Sun is* $9.3x10^{7}$ *miles*
- *average responsivity as low as 6.2pA/K*

Note that these measured quantities (e.g., 6.2pA/K) are typically associated with a specific *measured property* (e.g., average responsivity). We studied ways in which to extract these kinds of *measured information* from documents.[1] The mining of such information is critically important across many domains – especially those involving search and exploration of scientific and technical articles. For instance, an optics researcher may wish to know whether the performance of Nd:YAG laser-pumped KTP parametric oscillators has ever been tested at wavelengths longer than 2.4μm. Full-text search engines using inverted indexes allow *ad hoc* queries on terms such as "KTP parametric oscillator," but the ability to further filter search results based on wavelengths greater than 2.4μm is not typically supported. To accomplish this, one must first identify and extract valid *measured quantities* (e.g., 2.4μm) in unstructured text and then identify and extract the *properties* being measured (e.g., *wavelength*). These extractions

[1] We define measured information as measured quantities and the measured properties with which they are associated.

could then be stored in the index of a search engine in a way that supports subsequent document queries on measured information (e.g., faceted navigation, numeric range queries).

Surprisingly, there is very little existing work on how best to realize this process. Indeed, numerous challenges exist. For instance, there is a great deal of heterogeneity in how *measured quantities* and *measured properties* appear in text – both naturally and through corruption resulting from text conversion (e.g., converting a PDF to plain text). This, then, motivates the current investigation of how best to extract such information. Our contributions are as follows:

- We propose and describe a rule-based entity extractor to identify *measured quantities* in unstructured text documents. Our method includes an error-correcting procedure that recovers from aforementioned text conversion errors by 1) *reverse engineering* the corrupted and mangled measured quantities back to their original, correct form and 2) *standardizing* this form for storage in an inverted index and subsequent query processing.
- Using these extracted measured quantities, we show how to further extract the *measured properties* with which they are associated.
- Finally, we present MQSEARCH: the realization of a search engine with full support for *measured information*. MQSEARCH is a facet-based navigation system that allows users to navigate large document sets based on measured quantities, measured properties, and the topics and themes with which they are associated. To the best of our knowledge, no other search engine in existence fully supports such a capability.

We begin by describing the extraction of *measured quantities*.

MEASURED QUANTITIES

We view *measured quantities* as a 5-tuple of the form: (*sign, number, error, scientific notation, units*), where underlined elements are mandatory and others are optional. As an example, a team of researchers in Italy recently reported the first direct measurement of gravity's curvature as (1.3999±0.003) $x10^{-5}s^{2}m^{-1}$ (Rosi et al. 2015). The corresponding 5-tuple representation of this[2] is

(<empty>, 1.3999, 0.003, 10-5, s-2m-1).

5-tuples such as this are populated using a series of extraction rules that operate on individual sentences. These rules fall into four broad categories: 1) pre-processing, 2) units, 3) quantities, and 4) post-processing. Simplified forms of some of the rules for units and quantities are shown in Table 1.[3] We refer to the algorithm implementing such rules as *Measured Quantity Extractor* or **MQE**. We begin with pre-processing rules.

[2] Since there is no explicit sign in this example, the first element is left empty.

[3] Rules are shown in Perl-like syntax, the *de facto* standard for regular expressions.

Table 1. [MQE Rules.] Simplified forms of some rules for extraction of measured quantities

Rule	Pattern	Example Matches
1) number	[+−]?(\d{(\d?\d?[,]\d{2,3}){[,]\d{2,3}}*}|\d*)(\.(\d[\d\s]*\d|\d))?	1000.05, +5, -0.2, and 1,000
2) number (leading point)	[+-]?\.\d(\d\d(\s\d{3})+(\s\d{1,3})?|\d*)	-.98, .04, +.755
3) error	(\s{0,2} ± \s{0,2}|\d.|+)?	±0.003 in "1.3999 ± 0.003"
4) sci. notation.	(\s*[eE]\s*([xX×])\s*10 *\^? [+-]?\d+)?	*e.g.*, forms of $\times 10^{-5}$: ×105, e-5, E-5
5) unit	*e.g.*, [fpnμmcdk]?m([\^]?[2-6] | [\-][1-6]) — $m^{\#}$ **normalized to** $m^{\wedge}\#$	*μm*, *m*−1 (m^{-1}), *cm*2 (cm^2), *cm*^2
6) connector	(\s?/\s? | [Pp]er [-per-| [-\s× ·*])?	per, /, ·, ×
7) compound unit	<unit>(<connector><unit>)+	km/h, kilometer per hour,$km{\cdot}h^{-1}$

Pre-Processing. As mentioned previously, when extracting text from various document formats (e.g., PDF, MS Office), characters often appear inconsistently. Minus signs, multiplication signs (e.g., x, ·), equal-like symbols (e.g., ≈, ≃, ≅), degree symbols, and the μ character can appear in a variety of ways or, in some cases, as "garbage" characters. For instance, minus can appear as the *en dash* character or appear corrupted as æ€. Pre-processing rules identify these variations in text and perform the necessary normalization for accurate extraction of units and quantities.

Units. *A measurement unit* preceded by a numeric string conforming to the 5-tuple structure is the base indicator of a measured quantity. Thus, to identify valid *measured quantities*, we constructed a comprehensive taxonomy of units from multiple public sources. Each unit has an associated rule. An example rule for *m* (i.e., symbol for meters) is shown in Rule 5 of Table 1. Note that such rules include optional prefixes for submultiples and multiples (e.g., μ before *m*, *kilo* before *meter*). Unit rules, when combined with pre-processing rules described previously, can accurately extract units under a range of noisy conditions. For instance, the corrupted unit *mâ€1* is correctly recovered as m^{-1} by MQE. Finally, as shown in Rules 6 and 7, compound units are also supported (e.g., km/h, kilometer per hour, $s^{-2} \cdot m^{-1}$).

Quantities. Like units, quantities (i.e., numbers with optional error ranges and scientific notation) can appear in a range of ways due to corruption and natural variation. These variations are collectively captured by rules such as those shown in Table 1 (i.e., Rules 1 through 4), which populate the remainder of the 5-tuple structure. As shown in Table 1, such rules capture a wide range of quantity formats (e.g., 10,000 with a comma, $1.3999 \pm 0.003 \times 10^{-5}$ with both an error range and scientific notation, 1.23x105 with lost exponent in 10^5).

Post-Processing. We have already seen that text extracted from various document formats can be noisy. For instance, information from tables, headers, and figures can sometimes result in seemingly random sequences of numbers and letters in extracted text. In some cases, such information can be picked up erroneously by aforementioned rules as *measured quantities*. This is especially true for single letter units (e.g., *m* for meters, *A* for ampere). Post-processing rules are employed to reject such extractions and minimize false positives. Examples of such rejection rules include context-based rules (e.g., reject when preceded by "Table" or "Figure"), repetition-based rules such as rejecting compound units consisting of repeated

single letter units (e.g., 3 AJmm), and allowing a dash only between certain quantities and units (e.g., *10-cm* is okay but not *10-A*).

MEASURED PROPERTIES

We now turn our attention to the extraction of *measured properties*. To better illustrate the problem, we show several example snippets containing *measured quantities*. In each example, the measured quantity is shown in blue, the *property* being measured is highlighted in red, and the characters connecting them are underlined:

- *a pixel pitch as high as roughly 352μm*
- *a 352μm pixel pitch*
- *The pixel pitch employed was 352μm.*
- *average gravity curvature ζ=(1.3999±0.003)x10^{-5}s^{-2}m^{-1}*
- with *50mL of 30% fuming sulfuric acid*
- *size* ≅ 0.1m^2
- *frequency of longitudinal scan was approximately 300 Hz*
- *a nominal current density of 1.3 A/cm^2 to 0.03 A/cm^2*
- *panel strength lower than 8.90 ksi (61.4 MPa)*
- *wavelengths at least 2.4μm*
- *large fields of about, or above 10 kV/cm*

From just the examples shown, it is easy to see that there is a high degree of variability in the words connecting a *measured property* with a *measured quantity*. However, upon closer inspection, we find that this variability can be reduced to a small number of syntactic patterns based on parts of speech (POS) that capture most scenarios. Table 2 shows some syntactic patterns that we employ to extract *measured properties*. We refer to the extractor applying such syntactic rules as *Measured Property Extractor* or **MPE**. The notation shown here employs the Penn Treebank format, which associates tags with POS (e.g., NP represents a noun phrase). In addition, the EQ tag represents all symbols related to '=' (e.g., ≈,≃), and the SYM tag matches one or two character symbols (e.g., a Greek letter).

EXPERIMENTAL EVALUATION

We evaluated our approach on a text corpus consisting of 40,807 unclassified technical reports published in the 2008–2010 time frame and hosted by the Defense Technical Information Center (DTIC). This rich collection describes a wide range of research funded by the DoD spanning numerous fields from engineering and physical science to biomedical research and social

Table 2. [MPE Rules.] Simplified forms of some syntactic patterns to extract measured properties

Pattern	Example Matches (two examples shown for each rule)	
NP SYM{0,2} EQ mq	1) *gravity curvature* $\zeta = 1.4 \times 10^{-5} s^{-2} m^{-1}$	2) *floor area* ≅ 32m^2
mq IN? NP	1) *a 352 μm pixel pitch*	2) *50mL of 30% fuming sulfuric acid*
NP IN DT? NP VP+ (TO|IN|RB|JJ)* mq	1) *strength of panel was set to 9 ksi*	2) *freq. of scans was roughly 300 Hz*
NP (IN DT? NP)* VP+ (IN|TO|RB|JJ)* mq	1) *pixel pitch employed was 352 μm.*	2) *panel strength was recorded at 9 ksi.*
NP (CC|IN|TO|RB|JJ)* \(?mq\)?	1) *wavelengths of at least 2.4 μm*	2) *panel strength (9 ksi)*

science. Table 3 shows the precision and recall estimates for both the MQE and the MPE over the entire corpus.

Table 3. 95% Confidence Intervals for precision and recall when extracting measured quantities (using MQE) and measured properties (using MPE) from the DTIC corpus.

Extractor	Precision	Recall
MQE	(0.93, 0.99)	(0.92, 0.99)
MPE	(0.93, 0.97)	(0.88, 0.94)

As can be seen in the table, both MQE and MPE perform extraordinarily well in extracting *measured quantities* and the *properties* they describe from documents across disparate fields. Having demonstrated the success with which *measured information* can be mined, we now demonstrate how these extractions can be exploited in novel search applications.

AN APPLICATION: MQSEARCH

MQSEARCH is a realization of a search engine with full support for *measured information*. MQSEARCH is implemented using Apache Solr,[4] which supports full-text search, faceted navigation, and numeric range queries. During the process of indexing and ingesting the DTIC document set into our search engine, we apply our extractors to encountered text and store both *measured quantities* and *measured properties* in the search engine index. In addition, the search engine performs keyphrase extraction on documents using the KERA algorithm (Maiya et al. 2013). Using Solr facet queries, extracted keyphrases can be used to produce a tag cloud for any subset of the document set. Figure 1 shows the faceted navigation panel of MQSEARCH, which allows users to filter documents based on discovered measurement units, quantity ranges, and measured properties. In Figure 1, the measurement unit *U/mL* is selected. We see that there are 153 documents (out of roughly 40,000) mentioning this unit with quantities ranging from 0.001 U/mL to 10,000 U/mL. The property most frequently measured in *U/mL* is *penicillin*. From the tag cloud, we see that documents containing quantities measured in U/mL tend to

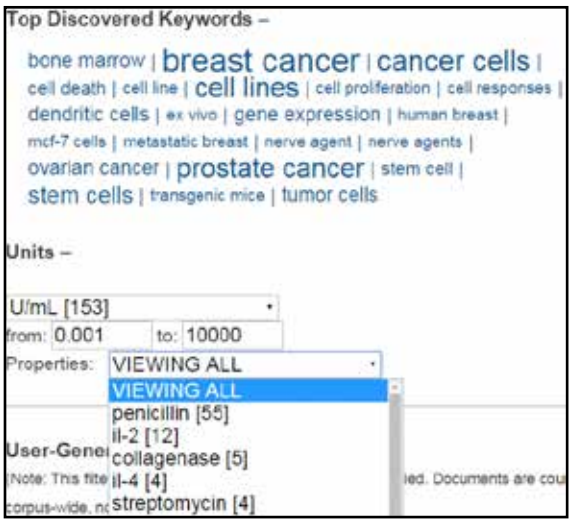


Figure 1. [MQSEARCH.] The measurement unit U/mL is selected, which reveals the associated topics (e.g., breast/prostate cancer), associated measured properties (e.g., concentrations of penicillin), and associated quantity ranges (i.e., 0.001 to 10,000)

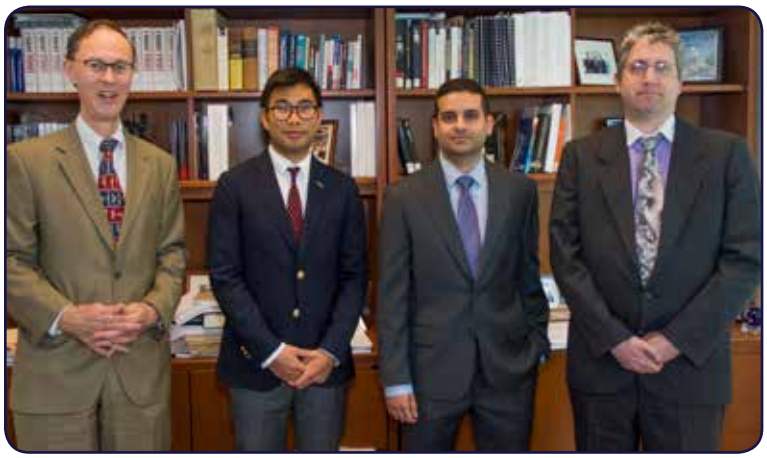
[4] http://lucene.apache.org/solr/.

cover topics such as breast cancer and prostate cancer research. The search results can be filtered further along any of these dimensions. To the best of our knowledge, ours is the first search engine with such support for *measured information*.

Dr. Maiya is a Research Staff Member in IDA's Information Technology and Systems Division. He holds a Doctor of Philosophy in computer science from the University of Illinois at Chicago.

Dr. Visser is a Research Staff Member in IDA's Information Technology and Systems Division. He holds a Doctor of Philosophy in physics from Yale University.

Dr. Wan is an Adjunct Research Staff Member in IDA's Information Technology and Systems Division. He holds a Doctor of Philosophy in computer science from Columbia University.



IDA President Dr. David S.C. Chu with the winners of this year's Welch Award: Dr. Wan, Dr. Maiya, and Dr. Visser.

The original article was published in the *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, August 2015.

Mining Measured Information from Text

http://dl.acm.org/citation.cfm?id=2766462&picked=prox

REFERENCES

Maiya, Arun S., John P. Thompson, Francisco L. Loaiza-Lemos, and Robert M. Rolfe. 2013. "Exploratory Analysis of Highly Heterogeneous Document Collections." *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD* '13, 1375–1383. New York, NY: ACM.

Rosi, G., L. Cacciapuoti, M. Menchetti, M. Prevedelli, and G. M. Tino. 2015. "Measurement of the Gravity-Field Curvature by Atom Interferometry." *Physical Review Letters* (ACM) 114 (1): 013001–013001-5.