On Establishing Behaviorally Adoptive Semantic Narratives

Jason Bryant, Gregory Hasseler, Matthiew Paulini
Air Force Research Laboratory/RI
Rome, NY 13411
{Jason.Bryant8, Gregory.Hesseler, Mathiew.Paulini}
@us.af.mil

Noor Ahmed
Dept. of Computer Science
Purdue University and AFRL/RI
West Lafayette, IN 4906
ahmed24@purdue.edu

***Abstract*— Providing personalized consumer contents can be both empowered and simplified through adapting analytics modeling and results with semantic representations. Analytics representations tend to be unique to their proprietary technological solutions, growing silos of non-interoperable, non-shareable results. Our approach to overcoming these obstacles is to pair our analytical modeling solution, Direct Qualification, with middleware integrated algorithms for graph, content, identity, and behavioral-based analytics. This abstraction layer of semantically represented analytics enabled multiple best-fit analytics engines to be deployed in parallel while providing a common query front-end for analytics observations, provenance, and trends. We introduce the establishment and the adaptation of a Behavior Ontology (BO) and Behavior Analytics (BA) modeling. We describe the integration of such behavior modeling with the semantic modeling of analytics and state management for an effective consumer content personalization system. We illustrate our prototype with publish and subscribe middleware and show the preliminary results. These components will be integrated into a holistic semantic analytics solution with autonomous functions for behavior optimizations, pluggable algorithm components, and end-to-end, machine learned personalization for information consumers and producers.**

Keywords-semantic modeling; consumer behavioral modeling; direct qualification; transactional pattern matching.

I. INTRODUCTION

Modern information management systems perform an increasingly expansive catalog of operations with greater complexity and with more feature expectations, including personalized consumption of information that requires analytics-based solutions. The desired feature gain from these efforts include more in depth data analysis, business processing autonomy, provenance traceability, trusted information sharing, and enhanced query options, all while optimizing performance. The current landscape of analytics generates a multitude of unintended negative consequences. Analytics engines are generally not interoperable, resulting in multiple silos of analytic results which cannot be simply joined, queried, or introspected for quality. Analytics engines rely on multiple proprietary standards with completely different paradigms of information, including content parsing, graph traversals, rule engine deductions, keyword or vector modeling for frequency, image & video learning, etc. Additionally, analytic technologies tend to take one out of three common approaches, including deductive, inductive, or behavioral. Our effort seeks to empower middleware solutions by combining all three approaches into a single queryable semantic service.

Modern information middleware consists of data models (Deductive Capabilities) and data analysis (Inductive Capabilities), with a limited notion of identity management for authorization and authentication, and services that orchestrate these capabilities according to consumer needs. Generally, OWL and RDF solutions approach problems within the web domain, which is distinct from an Information Management (IM) system in several ways. Search engines and web sites have autonomous feedback mechanisms that can capture rough estimations of information quality by observation of consumer link selection. Alternatively, IM systems have a more difficult time assessing quality from solicited consumers due to a lack of knowledge about information interactions once results have been returned, however IM systems have other advantages. Unlike semantic web-domain solutions, IM systems have access to identity management provenance, information analytics, information sourcing, as well as broad information access across multiple information dimensionality, including roles, formats, types, and access to deployed workflow or process models.

In our approach we seek to develop a middleware IM system that leverages the internal model, service, identity management, and data components to make more advanced information analytics and personalization possible. The effects of these enhancements can be positive facilitators for IM system, participant, and information trust, as well as assessments that can score and compare information quality, information impact, algorithm effectiveness, or model / ontology quality.

A key part of our approach is to seed the information process within the IM system with several algorithms that, when integrated, offer critical capabilities to autonomously learn information domains (topic modeling), personal information preferences (affinity clustering), and are informed by models prescribing the general workflows of participating information roles. In this manner the application narratives can be established

ISBN: 978-1-61208-420-6

and adapted through the dynamic topic models, the consumer behaviors can be observed and analyzed partially through models (behavior ontology) and partially through data-driven induction (affinity clustering), while all values and analytic results are represented via OWL and RDF, providing a huge advantage over existing middleware systems that require complex orchestrations or combinational queries that span multiple deployed analytics engines.

The middleware system has some modeling and infrastructure components complete, such as the analytics representation into OWL and RDF (direct qualification), and a set of pre-loaded algorithms (VSM, PageRank, HITS, topic modeling), others are still under development, including the behavioral ontology and the collaborative filtering algorithm that blends the consumer behaviors with feedback from the affinity scoring. Our main contributions can be summarized as follows:

- We outline a holistic semantic system that can support the abstraction of multiple analytics engine, eliminating concerns for proprietary analytics silos, complex combinational queries, and incompatible result representation.
- We introduce a simple consumer behavior ontology model that enables semantic representations of IM system transactions.
- Efficient scheme of integrating consumer behavior models, semantic modeling, and transactional pattern matching for content personalization.
- We apply semantic web applications to IM system domain with distinct information and modeling challenges.
- We project multiple analytic paradigms into a common semantic representation to enable more powerful queries and analytical tooling.

We have organized the paper as flows: we first give a brief background of the subject and the motivation behind our work in section II. Our consumer behavior modeling techniques and planning are discussed in section III, followed by the adaptation and the integration of these models into the semantic modeling and analytics in section IV. We discuss our prototype design and implementation in section V, and the preliminary experiments in section VI. In section VII and VIII covers the related work and the conclusion respectively.

II. BACKGROUND

The essence of information personalization is to enable the pairing of predictive analytics within adaptive content capabilities. Accomplishing this within an Information Management Systems requires following a foundational capabilities. These include:

- OWL Representation of Analytic Provenance and Results (Direct Qualification).
- Prescribed Workflow Activity Models (Constrained by Behavior Ontology & Planned Machine Learning Component).
- OWL Representation of Transactional Behavior History (Behavior Ontology)
- IM System Supported Analytics Engines (Jung, Lucene, and Mallet)
- Dynamic Application/Consumer Narrative Algorithm (Topic Modeling via Mallet)
- Personalized Query Results Feedback via Data Affinity Algorithms (Affinity Cluster Scoring via Jung’s PageRank and Lucene’s VSM)

Supporting the semantic modeling and execution of analytics, we created approaches including Direct Qualification (DQ) and State Management (SM) for OWL in our previous work [1]. In this work, we seek to extend these models to support the introspection of behaviors and interactions between consumers and the IM system (middleware), and thereby measure and validate the effectiveness of the behavioral analytic approaches and models themselves.

By embracing a data perspective that combines relationships for unstructured knowledge representation with structured, document-centric relationships, the process of determining, modeling, and expressing personalized information relevance with semantic technologies can be performed. Our approach seeks to solve a combination of challenges within Information Management (IM), Semantic Information Modeling, Data to Information (D2I), and Quality of Service (QoS) Enabled Dissemination (QED).

III. BEHAVIORAL ONTOLOGY AND ANALYTICS MODELING

Consumer and producer behaviors leave fingerprints at the informational layer that can be discovered, tracked, analyzed, correlated, and mined. Analytics engine complex event processing can utilize behavioral metadata, information content, analytics results, and historical trends to correlate information with emerging common consumer narratives over time. Additionally, linking these narratives to data-driven analytics assessment can monitor narrative changes, behavioral anomalies, determine critical personalized information, adapt to trends, and identify where information may be insufficient for consumer needs.

Behavior will be a distinctly different set of transactions, dependent upon the system upon which it is modeled. For our efforts, the IM system is a RESTful system allowing producer publications with attached metadata tags for type, format, and identity, and consumers that submit XPATH, SPARQL, or keyword queries. For a simplistic transactional IM system like this, we have started with a similarly simplistic behavioral ontology consisting of entities for Consumer, Producer, Query, QueryExpression, Publication, QueryType, ResultSet, Result, Document, Type, Format, Role, and Identity. The object properties and data properties associated are simply possessives of these entities (e.g. hasResultSet, performedQuery, publishedDocument, etc.). Capturing these relationships over time provides URIs that can act as hooks for analytic results, such as affinity scores across identity, format, type, document, or consumer dimensions.

The general process by which behaviors are associated with either published or consumed information is illustrated in Figure 1. After behaviors are collected within transactional provenance, either the published document or set of results is scored for affinity and related to consumer/producer identities.

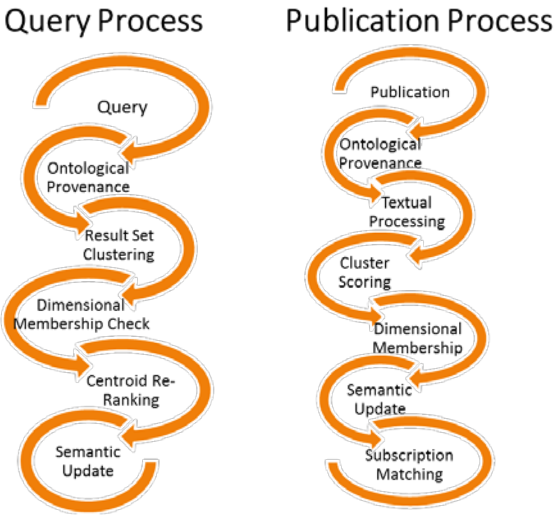

Figure 1- Process Overview

IV. SEMANTIC NARRATIVE INTEGRATION

Creating topic models, effectively producing dynamic information domain ontologies, on the fly is effective only if done utilizing sound techniques and dictionary to ontology classifications. WordNet and Mallet have built in tools that perform these features adequately, utilizing LDA to produce dynamic lists of topic models.

In our approach topic models are filtered according to configured quality metrics, and then assigned dimensional information categories. After being categorized as a topic model, the 2D grid of analytic scores for single dimensions such as role, query purpose, format, geography, or temporality. These can be subdivided into a 3-dimensional information centroid by utilizing cross-cutting attribution and grouping, most effectively according to identity.

Utilizing identity, role, and tasking metadata attributes to track information system transactions can support real-time content adaptation to information consumers by measuring trust and metadata-based information hotspots. Topic models cataloged via analytics can establish on the fly information domain ontologies, and paired with semantically modeled analytics, can result in advanced combinational forms of information and analytics queries. These customized data services enable topically modeled semantic narratives, behavioral adaptation, and content personalization via low-cost, reality-based solutions, rather than high-cost, prescribed, model-based solutions. Mallet is leveraged to perform LDA upon each received document, and when a minimally necessary set is received within an information dimension (format, type, role, identity, activity, etc.), a set of 30-40 dynamic topics are extracted across each identity, and then re-oriented around new centroids over time as the personalization features adapt.

A. STATE MANAGEMENT

Maintaining a historical record by creating a new instance of a resource and its present relationship states can generate mass duplication and waste memory and processing resources. Over time, particularly if an event has relationships that change quite often, numerous of versioned instances of the same event could be created, making queries overly complex and resulting in a high degree of overhead due to duplication for relationships may or may not be static.

Some attributes of an asset may be occurrent (e.g. name, identity, asset type, etc.), while others are continuant (e.g. fuel level, latitude, longitude, role, etc.). Semantics, even when using instances of an entity, treat all relationships as occurrent, although there is allowance for limiting their cardinality. OWL, SPARQL, and most ontologies do not have a built in mechanism to support the distinction between occurrent and continuant relationships. In order to retrieve changes of state for a data or object attribute of an instance, that attribute must be explicitly defined within an ontology or an additional, customized layer of abstraction.

Traditional semantic data model approaches fall short when confronting the challenge of state-based relationships. They focus on static knowledge representation, extractions of static data properties, or enabling of information management features via rule engines and inferencing. Managing states for data and object properties are applicable to all stateful semantic resources. Managing the state of semantic relationships is significant in reducing the computation time of semantic queries, the load on semantic DBs, and eliminating wasteful property and instance duplications. The key relationship used for this is the `specializationOf` predicate of the Provenance Ontology (Prov-O) W3C recommendation. It is intended to apply state-based relationships to any Entity, Agent, or Activity, auspices under which any semantic asset instance should fall.

In our experiments we apply this approach by requiring all occurrent relationships to be related to the singular instance URI of a specific entity, while all relationships involving state changes are related through Specialization deltas, such as a Consumer entity having a temporary role relationship.

B. DIRECT QUALIFICATION

Reification, an intrinsic complexity of the semantic standards, is the consequence of attempting to simplify all relationships into Subject-Predicate-Object sets. It is normally implemented when a semantically modeled instance is seeking to express either the qualification or provenance of a relationship. These two cases can be mitigated without resorting to reification, however. Adopting a quad-based perspective of semantic relationships can achieve a basic form of provenance by allowing traceability to the source named graph's unique URI. The Prov-O ontology expands the set of provenance support and supplies some generalized predicates for qualification. This effectively solves the non-probabilistic subset of analytics use cases. However, even with pairing both of these approaches there is a failure to solve the qualification of probabilistic analytics, such as the results of Vector Space Modeling, PageRank, HITS, or other Natural Language Processing (NLP) analytics. Our approach towards supplying these capabilities is the Direct Qualification of probabilistic relationships with a supporting relevancy ontology.

The primary steps for enabling Direct Qualification are outlined as follows:

1. Support persistence for raw documents and semantic quad-based relationships, ideally by using the semantic URI of the document's named graph as the unique key for the raw document retrieval.
2. Strictly enforce the separation of the semantic models for class instances from the events affecting their state relationships.
3. Support graph-based processing of analytics over semantic edges and vertices.
4. Support event-based scoring triggers for analytics, such as SPARQL queries, XPath queries, semantic reasoning, or keyword searches of raw text.
5. Determine the appropriate Direct Qualification Model based upon tests for occurrence, continuance, and the monotonicity of the entities involved in the applied analytic.
6. Express the scoring of documents through the pairing of a provenance ontology with an analytics/relevancy ontology.
7. Persist the DQ results within the quad-store.

An example of DQ is illustrated in the following figures, with a general purpose (Figure 2) set of analytics relationships from the ontology, followed by more concrete examples utilizing PageRank (Figure 3) and VSM (Figure 4).

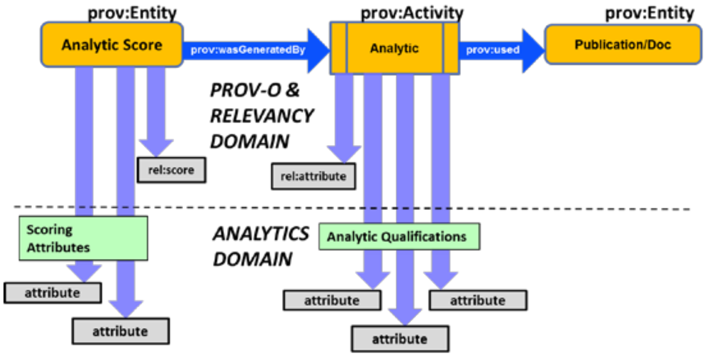

Figure 2 – General Direct Qualification Application

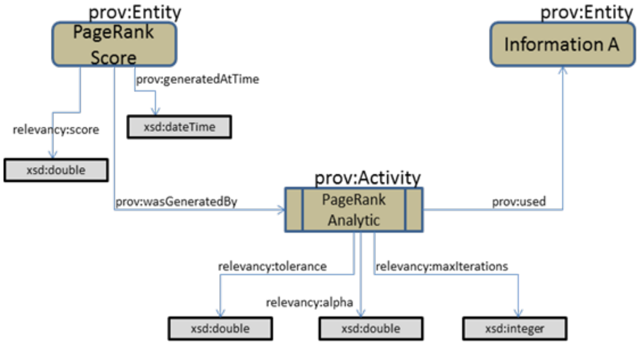

Figure 3 – PageRank Direct Qualification Use Case

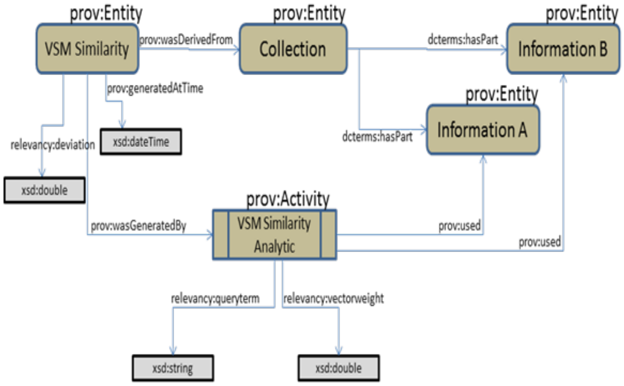

Figure 4 – VSM Direct Qualification Use Case

V. SYSTEM DESIGN AND IMPLEMENTATION

After stages for pre-processing, format determination (Aperture) and semantic extraction completes, the execution of the analytics, DQ, narrative creation, analytic scoring, and affinity scoring are executed as part of the publication and query processes. The system utilizes the following:

- Pre-requisite: Establishing a common representation for high level application narratives, research provenance, and analytic domain concepts.
- Pre-requisite: Establishing an ontology with entities and relationships supporting affinity, clustering, topic modeling, and role, format, type and identity-oriented membership groups.
- Pre-requisite: Create topic model classifiers for each primary information dimension of the metadata tags, including behavior, role, information format, geolocation, and identity.

1. Score each new publication according to a similarity / affinity vector within each primary information dimension.
2. Recalculate cluster centroids and dynamic topic model relations after every n publications.
3. Evaluate thresholds for information grouping inclusion / exclusion when thresholds of affinity are reached for each measured information dimension.
4. Score the information for aggregate personalization for cross-information domain grouping, according to existing queries.
5. Adapt the baseline metric for personalized relevancy thresholds after an initial n publications.

- Post-Operations: Measure the transition of information centroids / clusters over time as classifiers improve.
- Post-Operations: Compare results of the trend-based relevancy metric to a prescribed workflow template in order to validate models.

The functional composition of the system is illustrated below (Figure 5) by showing the internal resources and capabilities established and utilized in order to bridge the gap between raw information, behavior, validated narratives, and personalized content.

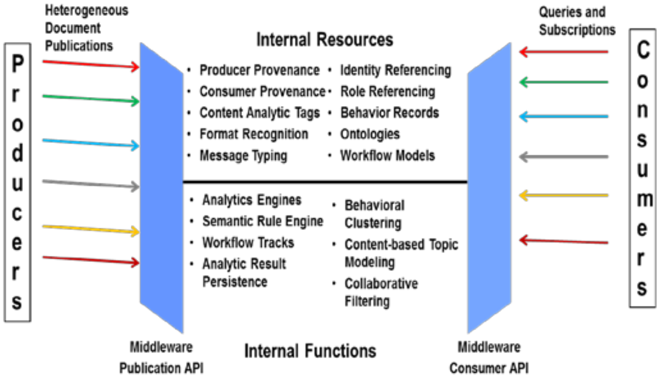

Figure 5 – Mapping of Middleware Resources/Functions

VI. Experiments

The test scenarios include a compilation of 10000+ research papers of disparate fields and a 15 GB set of imagery from Flickr. The semantic relationships created were produced by means of the extraction framework we created in our previous ICCRTS research (Bryant, 2014), with added support for GeoSPARQL location extractions.

The results of semantic processing is an semantic document represented via RDF/OWL relating internal values for details involving times, locations, narratives, cataloged topic models, points of contact, metadata, etc.

Ontology support includes common solutions for time, geospatial (GeoSPARQL), common elements (U-Core SL), and custom ontologies for information management, and relevancy. Format and XML type determination is performed in a pre-processing stage prior to semantic extraction, if applicable. The Aperture source project is adopted to provide the majority of the format and type determination solution.

The results of the experiments include semantically represented cross-dimensional domain membership of all published data, according to the determined applicable information domains, and the centroid k-means scores of each information and identity-based data dimension. This enables the determination of domain-based relevancy measures, and in particular, cross-cutting identity-oriented information relevancy measures.

VII. Related work

This research addresses trust, affinity, collaborative filtering, and cross-cutting analytics engine abstractions for interoperable queries. This supports the modeling and probabilistic analytics that assesses and qualitatively relates information, enhances query options, and detect information anomalies or model workflow outliers. These capabilities provides potentially critical advancements towards autonomously determining whether information should be excluded from an information result set based upon its determined value, historical precedence, and personalized interest. While this allows queries to alter from a pre-defined domain-based set of operations, it also eliminates extensive modeling and domain ontology costs.

As semantic standards mature and applications expand into new domains, research regarding semantic management of stateful relationships is beginning to be explored more fully. Current research has been tangential, at best, while missing many of the niche problem areas of semantics. Approaches in this area have focused on inferencing through the use of join sequences [5] or resolving models with conflicting states [4]. So far, approaches involving applied analytics for state management have attempted to do so during the extraction phase of data [6], rather than utilizing semantic technologies or ontology models.

ISBN: 978-1-61208-420-6

While there are ongoing efforts towards document analysis using analytics such as PageRank (Ding, 2002), VSM, and HITS, the focus of those efforts has been on ontology matching [2] or temporal/geospatial query enhancement [8]. Our approach differs in that it stays confined to semantic technologies with special emphasis on event-based information sharing, modeling, data mining, and retrieval, all combined. Furthermore, one of the key differentiator of some existing semantic models with DQ approach is that they adopt a constantly "present" based view that updates the instance with relationships reflecting any changes in its state. Thereby, the state changes' value can never be considered truly distinct from its identity URI.

While some approaches were discovered that sought to model analytics similar to the direct qualification and semantic state management techniques, none were found that sought to provide OWL and RDF abstractions and system support to facilitate analytic engine interoperability and freedom from multiple proprietary query dependencies.

VIII. CONCLUSION

Applying these models, ontologies and approaches to a new type of information set can make that information, and its relevancy score results, more discoverable and of higher quality. The most critical takeaway from this work is that the advantages of semantic standards and reasoning can be leveraged upon analytic provenance and results, providing a common query representation, eliminating the need for proprietary or complex combinational queries that span multiple analytical data silos. Also critically, behavioral observations expressed with DQ can be leveraged with dynamically adaptive consumer usage narratives to build powerful semantic functionality that augments traditional SPARQL queries for simplistic data extractions that were ignorant of behavior, analytics, or consumer information affinities, with new features that, essentially, enable an autonomous collaborative filtering process that is represented 100% via semantic standards.

Autonomous collaborative filtering would itself be a powerful feature, but leveraged with semantic technologies that bridge extensible pluggable ontologies, while simultaneously abstracting analytics-engine queries and personalizing information to consumer needs, could enable novel research, new information and web functionality, and act as a unifying analytics front-end.

In our future work, we will explore semantic state traceability paired with diverse analytics. Reasoning over stateful trends within segmented time periods can demonstrate possible advanced uses of semantics for stochastic, graph-based, boolean-based, or other analytics algorithms, thus producing support for personalized prioritization, query result set ordering, and provenance modeling of analytics. Additionally, the enhancements from this work could enable determinations of efficiency for different analytics, and have the potential to combine analytic-based queries with semantic queries.

REFERENCEs

[1] Bryant, J., Paulini, M., Hasseler, G., Lebo, T., "Enhancing Information Awareness through Directed Qualification of Semantic Relevance Scoring Operations", ICCRTS, 2014

[2] Tous, R., Delgado, J., "A Vector Space Model for Semantic Similarity Calculation and OWL", IN DEXA, 2006.

[3] Ding, C., He, X., Husbands, P., Zha, H., Horst, D., "PageRank, HITS, and a Unified Framework for Link Analysis", 25th SIGIR Proceedings, 2002

[4] Zhang T, Xu D, Chen J. Application-oriented purely semantic precision and recall for ontology mapping evaluation. Knowl-Based Syst 2009;21(8):794–799.

[5] Kai Zeng , Jiacheng Yang , Haixun Wang , Bin Shao , Zhongyuan Wang, A distributed graph engine for web scale RDF data, Proceedings of the VLDB Endowment, v.6 n.4, p.265-276, February 2013.

[6] P. D. Turney and P. Pantel, "From Frequency to Meaning: Vector Space Models of Semantics", Volume 37, pages 141-188, 2010.

[7] Das, SR. Bunescu, R. Mooney. "Collective Information Extraction with Relational Markov Networks". 42nd Annual Meeting of the Association for Computational Linguistics, July, 2004

[8] Perry, M., Sheth, A., Arpinar, I., "Geospatial and Temporal Semantic Analytics", Encyclopedia of Geoinformatics, 2009.

[9] Bryant, J., Paulini, M., "Making Semantic Information Work Effectively for Degraded Environments", ICCRTS, 2013.

[10] K. Sudo, S. Sekine, R. Grishman. "An Improved Extraction Patterns Representation Model for Automatic IE Pattern Acquisition". 41st Annual Meeting of the Association for Computational Linguistics, July, 2003.

[11] Lebo, T., Graves, A., McGuinness, D., "Content-Preserving Graphics", Consuming Linked Data Conference, 2013.

[12] J. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel, M. Tyson. "FASTUS: A Cascaded Finite-state Transducer for Extracting Information for Natural-Language Text". Finite-State Language Processing. MIT Press, Cambridge, MA. 1997

ISBN: 978-1-61208-420-6