The Contextual Features in Schema-Agnostic Environment

Muhammad Ali Norozi
mnoroz@idi.ntnu.no

Department of Computer and Information Science
Norwegian University of Science and Technology

04th April 2014
Trondheim, Norway
Outline

1 Introduction and Motivations
2 Background and Challenges
3 Research Results
4 Experimental Evaluations
5 Conclusions
Outline

1 Introduction and Motivations
2 Background and Challenges
3 Research Results
4 Experimental Evaluations
5 Conclusions
Semi-structured and/or Schema-agnostic data

Semi-structured and schema-agnostic data $D \subseteq$ of:

- neither raw data nor strictly typed.
- “schema-less” or “self describing”.
- schema is known and operationalized by the retrieval system but unfamiliar and/or unknown to the user.
- the ability to accommodate variations in data.
- flexible, full-text (keyword) +/- structured query support.
Semi-structured and/or Schema-agnostic data

Semi-structured and schema-agnostic data $D \subseteq$ of:

- neither raw data nor strictly typed.
- “schema-less” or “self describing”.
- schema is known and operationalized by the retrieval system but unfamiliar and/or unknown to the user.
- the ability to accommodate variations in data.
- flexible, full-text (keyword) +/- structured query support.
Semi-structured and/or Schema-agnostic data

Semi-structured and schema-agnostic data $D \subseteq$ of:

- neither raw data nor strictly typed.
- “schema-less” or “self describing”.
- schema is known and operationalized by the retrieval system but unfamiliar and/or unknown to the user.
- the ability to accommodate variations in data.
- flexible, full-text (keyword) +/- structured query support.
Semi-structured and/or Schema-agnostic data

Semi-structured and schema-agnostic data $D \subseteq$ of:

- neither raw data nor strictly typed.
- “schema-less” or “self describing”.
- schema is known and operationalized by the retrieval system but unfamiliar and/or unknown to the user.
  - the ability to accommodate variations in data.
  - flexible, full-text (keyword) +/- structured query support.
Semi-structured and/or Schema-agnostic data

Semi-structured and schema-agnostic data $D \subseteq$ of:

- neither raw data nor strictly typed.
- “schema-less” or “self describing”.
- schema is known and operationalized by the retrieval system but unfamiliar and/or unknown to the user.
- the ability to accommodate variations in data.
- flexible, full-text (keyword) +/- structured query support.
Semi-structured and/or Schema-agnostic data

Semi-structured and schema-agnostic data $D \subseteq$ of:

- neither raw data nor strictly typed.
- “schema-less” or “self describing”.
- schema is known and operationalized by the retrieval system but unfamiliar and/or unknown to the user.
- the ability to accommodate variations in data.
- flexible, full-text (keyword) +/- structured query support.
Relevancy in Schema-agnostic environment

Relevancy

Relevancy in schema-agnostic environment should:

- Take advantage of the “semantic” structure of the document.
- Return most specific part of the document instead of whole document.
- Effects of the context or more specifically structural context in retrieval.

- Should effectively calculate the weights on the logical documents in reasonable time.
Relevancy in Schema-agnostic environment

Relevancy

Relevancy in schema-agnostic environment should:

- Take advantage of the “semantic” structure of the document.
- Return most specific part of the document instead of whole document.
- Effects of the context or more specifically structural context in retrieval.
- Should effectively calculate the weights on the logical documents in reasonable time.
Relevancy in Schema-agnostic environment

Relevancy

Relevancy in schema-agnostic environment should:

- Take advantage of the “semantic” structure of the document.
- Return most specific part of the document instead of whole document.
- Effects of the context or more specifically *structural context* in retrieval.
- Should *effectively* calculate the weights on the logical documents in reasonable time.
Relevancy in Schema-agnostic environment

Relevancy

Relevancy in schema-agnostic environment should:

- Take advantage of the “semantic” structure of the document.
- Return most specific part of the document instead of whole document.
- Effects of the context or more specifically structural context in retrieval.

- Should effectively calculate the weights on the logical documents in reasonable time.
Relevancy in Schema-agnostic environment

Relevancy

Relevancy in schema-agnostic environment should:

- Take advantage of the “semantic” structure of the document.
- Return most specific part of the document instead of whole document.
- Effects of the context or more specifically *structural context* in retrieval.

- Should **effectively** calculate the weights on the logical documents in reasonable time.
Example

Figure : Querying the search engine with full text.
Example

[PDF] Relevancy in **Schema Agnostic Environment** - IEEE-TCDL
by MA Noroz - Related articles
Relevancy in **Schema Agnostic Environment**. Muhammad Ali Noroz. Department of Computer and Information Science. Norwegian University of Science and ...

**What is database-agnostic?** - Definition from WhatIs.com
search.datamangement.techtarget.com/.../database-agnostic
by Margaret Rouse - in 1,586 Google+ circles
**Database-agnostic** software can be useful in an **environment** where data must be sourced from heterogeneous databases. On the other hand, because a ...

**Usability of Keyword-Driven Schema-Agnostic Search** - Springer
link.springer.com/chapter/10.1007%2F978-3-642-13489-0_24
by T Tran - 2010 - Cited by 11 - Related articles
The increasing amount of data on the Web bears potential for addressing complex information needs more effectively. Instead of keyword search and browsing ...
Example

Figure: Querying the search engine with NEXI query.
Example

Usability of Keyword-Driven Schema-Agnostic Search - Springer
link.springer.com/chapter/10.1007%2F978-3-642-13489-0_24 by T Tran - 2010 - Cited by 11 - Related articles
The increasing amount of data on the Web bears potential for addressing complex information needs more effectively. Instead of keyword search and browsing ...

[PDF] Relevancy in Schema Agnostic Environment - IEEE-TCDL
www.ieee-tcdl.org/Bulletin/v8n1/papers/norozif.pdf by MA Norozi - Related articles
Permission to make digital or hard copies of all or part of this work for personal or ... raries) and web search, schema agnostic search environment must also define a ... important than keywords lying in the title. 3. STATE OF ..... erature review and getting a field overview. It is expected to .... Journal of the American Society for.

Tutorial: First Steps with XMLBeans
xmlbeans.apache.org > xmlbeans.apache.org
An environment set up according to what’s described in Installing XMLBeans. ... You’ll create a few files as part of this tutorial. ... A schema-agnostic way to traverse the full XML infoSet. .... This XML represents a purchase order in through which Gladys, an amateur astronomer, is buying two books to use as references in her ...
Example

Usability of Keyword-Driven Schema-Agnostic Search - Springer
link.springer.com/chapter/10.1007%2F978-3-642-13489-0_24
by T Tran - 2010 - Cited by 11 - Related articles
The increasing amount of data on the Web bears potential for addressing complex information needs more effectively. Instead of keyword search and browsing ...

[PDF] Relevancy in Schema Agnostic Environment - IEEE-TCDL
by MA Norozi - Related articles
Permission to make digital or hard copies of all or part of this work for personal or ...
and web search, schema agnostic search environment must also define a ... important than keywords lying in the title. 3. STATE OF ..... erature review and getting a field overview. It is expected to .... Journal of the American Society for.

Tutorial: First Steps with XMLBeans
xmlbeans.apache.org > xmlbeans.apache.org
An environment set up according to what’s described in Installing XMLBeans. ... You’ll create a few files as part of this tutorial. ... A schema-agnostic way to traverse the full XML infoset. .... This XML represents a purchase order in through which Gladys, an amateur astronomer, is buying two books to use as references in her ...
Example

```xml
<article xmlns:xlink="http://www.w3.org/1999/xlink">
  <header>
    <title>Wiki markup</title>
    <id>42</id>
    <revision>
      <timestamp>2006-10-05 14:22</timestamp>
    </revision>
    <categories>
      <category>Markup languages</category>
    </categories>
  </header>
  <body>
    <section>
      <p>
        <b>Wiki markup</b> is used in
        <link xlink:href="../Wi/article2.xml" xlink:type="simple">Wikipedia</link>.
      </p>
    </section>
  </body>
</article>
```
Example

```xml
  <article>
    <header>
      <title></title>
      <id></id>
      <revision>
        <timestamp></timestamp>
      </revision>
      <categories>
        <category></category>
      </categories>
    </header>
    <body>
      <section>
        <p>
          <b></b>
          <link></link>.
        </p>
        ...
      </section>
      <section>
        <p>
          ...
        </p>
      </section>
      <section>
        ...
      </section>
    </body>
  </article>
```
Example
Example
Example
External and internal features

- Hyperlink structure in documents implicitly provides the “external” structural context of information.

- Internal / Hierarchical structure of documents provides “internal” structural context of information.

- These should be used as the contextual features surrounding the important and not so important information retrieved.
External and internal features

- Hyperlink structure in documents implicitly provides the “external” structural context of information.

- Internal / Hierarchical structure of documents provides “internal” structural context of information.

- These should be used as the *contextual features* surrounding the important and not so important information retrieved.
External and internal features

- Hyperlink structure in documents implicitly provides the “external” structural context of information.

- Internal / Hierarchical structure of documents provides “internal” structural context of information.

- These should be used as the *contextual features* surrounding the important and not so important information retrieved.
**Research Goal**

**RQ:** How to effectively use the structure within semi-structured documents as evidence in the pursuit of “further” improving retrieval effectiveness?
KEEP CALM
and formulate your RESEARCH QUESTION
Research Questions

**RQ1:** What is the role and the significance of the **structural context** in the ranking of the focused items, and what kind of structural context can be “beneficially” utilized?
**Research Questions**

**RQ2:** How can we improve the retrieval approaches which make use of the structural context, and subsequently, how should the retrieval effectiveness of those improved strategies be evaluated?
RQ3: How to improve the retrieval of small elements in focused retrieval?
Research Questions

RQ4: How can we effectively utilize the scoring of multiple systems to retrieve focused results at varied granularity levels with good-enough precision (scoring)?
Research Questions

RQ_{ef}: How *efficiently* can we carry out the semi-structured retrieval task?
Outline

1. Introduction and Motivations
2. Background and Challenges
3. Research Results
4. Experimental Evaluations
5. Conclusions
Semi-structured retrieval

Semi-structured data

(i) The structure is irregular — the same data might be organized in different ways.

(ii) The structure can be very large and at the same time constantly and rapidly evolving as well.
Internal Representation – hierarchical structure

Figure: Semi-structured data such as XML is usually represented as a Tree structure, and encoded using Dewey encoding or others.
Internal Representation – functions

\[
\begin{align*}
parent(n_j) &= n_i : (n_i, n_j) \in \{\text{set of edges}\} \\
ancestors(n_i) &= \begin{cases} n_j : n_j \in parent(n_i) \cup \\
&\exists n_k \in parent(n_i) \\
&s.t. \ n_j \in \text{ancestors}(n_k) \end{cases} \\
children(n_i) &= \{n_j\} : \forall n_j \ (n_i, n_j) \in \{\text{set of edges}\} \\
descendants(n_i) &= \begin{cases} n_j : n_j \in \text{children}(n_i) \cup \\
&\exists n_k \in \text{children}(n_i) \\
&s.t. \ n_j \in \text{descendants}(n_k) \end{cases} \\
siblings(n_i) &= \begin{cases} n_k : (n_i, n_j) \in \{\text{set of edges}\} \cap \\
&\exists k (n_k, n_j) \in \{\text{set of edges}\} \\
&\forall k \neq i \end{cases} \\
\kinship(n_i) &= \begin{cases} n_j : (n_i, n_j) \in \{\text{set of edges}\} \cup \\
&\exists n_k \in \text{children}(n_i) \cup \\
&\exists n_k \in \text{ancestors}(n_i) \cup \\
&\exists n_k \in \text{descendants}(n_i) \cup \\
&\exists n_k \in \text{siblings}(n_i) \\
&s.t. \ n_j \in \kinship(n_k) \end{cases}
\end{align*}
\]
Internal Representation – structural context
Internal Representation – structural context
Internal Representation – structural context
Internal Representation – structural context
External Representation
External Representation – matrix

\[ A = \begin{pmatrix}
0 & 1 & 1 & \varepsilon & \varepsilon \\
\varepsilon & 0 & \varepsilon & \varepsilon & 1 \\
\varepsilon & 1 & 0 & \varepsilon & 0 \\
1 & 1 & 1 & 0 & \varepsilon \\
1 & \varepsilon & 0 & 1 & 0
\end{pmatrix} \]

\[ A_{ij} = \begin{cases}
1 & \text{if there is a link from document } d_i \text{ to } d_j \\
\varepsilon & \text{if there is no link from document } d_i \text{ to } d_j, \text{ while there is a link from document } d_j \text{ to } d_i, \\
0 & \text{otherwise}
\end{cases} \]
Challenges

When **structural features** are not appropriately handled!
Challenges

1. Overlap

The same content appears several times in the results because of the overlapping or nested structural components.

- FCFC selection.
- Selection fusion.
- Others...
Challenges

2. Focused-ness

- Relevant item at the right granularity levels.
- Subject to different use cases and user interface scenarios.
Challenges

3. **Size bias**

- Neither small nor big elements dominate the top result list.
- There should be a way to: to accumulate, combine and propagate the evidence into the relevance of retrievable items, from their structural context, to lessen the effects of “size bias”.
What is the role and significance of the structure in the retrieval of relevant information, in schema agnostic settings?
Accumulation of evidences from Structural context

Contextualization – to explore the effects of structural context.
Accumulation of evidences from Structural context

Contextualization – to explore the effects of structural context.

- Scant textual evidence makes matching of small text units, such as paragraphs, hard.
Accumulation of evidences from Structural context

Contextualization – to explore the effects of structural context.

- Scant textual evidence makes matching of small text units, such as paragraphs, hard.
- Retrievable units are not independent of each other, and typically appear in context of other retrievable units.
Accumulation of evidences from Structural context

Contextualization – to explore the effects of structural context.

- Scant textual evidence makes matching of small text units, such as paragraphs, hard.
- Retrievable units are not independent of each other, and typically appear in context of other retrievable units.
- Text passages semantically dependent on each other within a document based on the *Structure*.
Accumulation of evidences from Structural context

Contextualization – to explore the effects of structural context.

- Scant textual evidence makes matching of small text units, such as paragraphs, hard.
- Retrievable units are not independent of each other, and typically appear in context of other retrievable units.
- Text passages semantically dependent on each other within a document based on the Structure.

- **Contextualization** is a mechanism deviced to estimate the relevance of a given structural text or document unit with information obtainable from - besides the unit itself - the surrounding structural text or document units, i.e., from the “structural context” of the unit.
Contextualization

Exploring the features in the internal and external structural context of retrievable unit.

- Structural contextual features surrounding a document and element, contain traces of evidence, identifying the relevance of a document.

- Document in a lot of good neighbors form a good structural context, and should be used for retrieval effectiveness.

- Similarly elements in a good neighbourhood — ancestors, descendants, siblings, kinships form a good structural context, and should be used to boost elements’ retrieval effectiveness.
Figure: Hyperlink structure of five documents & context of P2.
Contextualized Doc

\( P_2 \)

Contextualizing Docs \( C_{P_2} \subseteq \{ \text{inlinks}(P_1, P_3, P_4), \text{outlinks}(P_5) \} \)

Contextualization Vector

\( \{ g(P_1), g(P_3), g(P_4), g(P_5) \} \)

\[
\begin{align*}
\text{CR}(P_2, f, C_{P_2}, g^k) &= BS(P_2) + f \cdot \left[ BS(P_1) \quad BS(P_3) \quad BS(P_4) \quad BS(P_5) \right] \cdot \left[ g(P_1) \quad g(P_3) \quad g(P_4) \quad g(P_5) \right] \\
&= BS(P_2) + f \cdot \left[ g(P_1) + g(P_3) + g(P_4) + g(P_5) \right] \\
&= \text{Basic Score} + \text{Magnitude} \cdot \text{Impact of Context} \\
&= \text{Context of } P_2
\end{align*}
\]
Contextualized Doc

\( P_2 \)

Contextualizing Docs \( C_{P_2} \subseteq \{ \text{inlinks}(P_1, P_3, P_4), \text{outlinks}(P_5) \} \)

Contextualization Vector

\[ \{ g(P_1), g(P_3), g(P_4), g(P_5) \} \]

\[ CR(P_2, f, C_{P_2}, g^k) = \underbrace{BS(P_2)}_{\text{Basic Score}} + \underbrace{f \cdot \begin{bmatrix} BS(P_1) & BS(P_3) & BS(P_4) & BS(P_5) \end{bmatrix}}_{\text{Magnitude}} \cdot \begin{bmatrix} g(P_1) & g(P_3) & g(P_4) & g(P_5) \end{bmatrix} \]

\[ \underbrace{g(P_1) + g(P_3) + g(P_4) + g(P_5)}_{\text{Impact of Context}} \]

Context of \( P_2 \)
Figure: Hierarchical Graph with context of element \(1.2.1.2\) (dewey encoding).
Contextualized El. $\langle 1.2.1.2 \rangle$

Contextualizing Els $C_{\langle 1.2.1.2 \rangle} \subseteq \{\text{ancestors}(\langle 1.2.1 \rangle, \langle 1.2 \rangle, \langle 1 \rangle)\}$

Contextualization Vector $\{g_{<1.2.1>}, g_{<1.2>}, g_{<1>}\}$

$CR(\langle 1.2.1.2 \rangle, f; C_{\langle 1.2.1.2 \rangle}, g^k) = BS_{<1.2.1.2>} + f \cdot \left[ BS_{<1.2.1>} \quad BS_{<1.2>} \quad BS_{<1>} \right] \cdot \left[ g_{<1.2.1>} \quad g_{<1.2>} \quad g_{<1>} \right]$

$= g_{<1.2.1>} + g_{<1.2>} + g_{<1>}$

Basic Score

Magnitude

Impact of Context

Context of $\langle 1.2.1.2 \rangle$
**Contextualized El.**  
\[ \langle 1.2.1.2 \rangle \]

**Contextualizing Els**  
\[ C_{\langle 1.2.1.2 \rangle} \subseteq \{ \text{ancestors}(\langle 1.2.1 \rangle, \langle 1.2 \rangle, \langle 1 \rangle) \} \]

**Contextualization Vector**  
\[ \{ g_{<1.2.1>}, g_{<1.2>}, g_{<1>} \} \]

\[
CR(\langle 1.2.1.2 \rangle, f, C_{\langle 1.2.1.2 \rangle}, g^k) = \underbrace{BS_{<1.2.1.2>}}_{\text{Basic Score}} + f \cdot \left\{ \begin{array}{c}
\begin{bmatrix}
BS_{<1.2.1>} & BS_{<1.2>} & BS_{<1>}
\end{bmatrix} \cdot \\
\begin{bmatrix}
ge_{<1.2.1>} & ge_{<1.2>} & ge_{<1>}
ge_{<1.2.1>} + ge_{<1.2>} + ge_{<1>}
\end{bmatrix}
\end{array} \right\}
\]

- **Magnitude**
- **Impact of Context**
- **Context of** \[ <1.2.1.2> \]
Contextualization model is re-scoring scheme
Good context (internal and external) is an evidence that is used here to deduce that the document / element is a good candidate to posed query.
Context materialization

Can the evidences lying loosely in the context, be intelligently materialized?
Can the evidences lying loosely in the context, be intelligently materialized?

- With the help of random walk on context and contextualization.
Can the evidences lying loosely in the context, be intelligently materialized?

- With the help of random walk on context and contextualization.
- Random walk independent and / or dependent of a query topic.
Generalized Formulation

\[ CR(x, f, C_x, g_{qi}^k, g_{qd}^k) = (1 - f) \cdot BS(x) + f \cdot \left( \frac{\sum_{y \in C_x} BS(y) \cdot g_{qi}^k(y)}{\sum_{y \in C_x} g_{qi}^k(y)} + \frac{\sum_{y \in C_x} BS(y) \cdot g_{qd}^k(y)}{\sum_{y \in C_x} g_{qd}^k(y)} \right) \]

Query Independent

Query Dependent

Context of \( x \)
### Structural Context

**CR\textsubscript{external}**
- Documents cited a lot, or documents containing more in-links form a good context?
  - \( C_x \subseteq \text{inlinks}(x) \)
- And what about outlinks.
  - \( C_x \subseteq (\text{inlinks}(x) \cup \text{outlinks}(x)) \)

**CR\textsubscript{internal}**
- Use internal hierarchical structure of the semi-structured documents (XML).
- XML Structure as graph.
- Myriads of random surfers.
- Which types of element context help to improve retrieval effectiveness?

**CR\textsubscript{hybrid}**
- Externally accumulated evidences re-enforce the evidence accumulated from within the hyperlinked and hierarchical XML.
Structural Context

$CR_{external}$
- Documents cited a lot, or documents containing more in-links form a good context?
  - $C_x \subseteq \text{inlinks}(x)$
- And what about outlinks.
  - $C_x \subseteq (\text{inlinks}(x) \cup \text{outlinks}(x))$

$CR_{internal}$
- Use internal hierarchical structure of the semi-structured documents (XML).
- XML Structure as graph.
- Myriads of random surfers.
- Which types of element context help to improve retrieval effectiveness?

$CR_{hybrid}$
- Externally accumulated evidences re-enforce the evidence accumulated from within the hyperlink and hierarchical XML.
Structural Context

$CR_{external}$
- Documents cited a lot, or documents containing more in-links form a good context?
  - $C_x \subseteq \text{inlinks}(x)$
- And what about outlinks.
  - $C_x \subseteq (\text{inlinks}(x) \cup \text{outlinks}(x))$

$CR_{internal}$
- Use internal hierarchical structure of the semi-structured documents (XML).
- XML Structure as graph.
- Myriads of random surfers.
- Which types of element context help to improve retrieval effectiveness?

$CR_{hybrid}$
- Externally accumulated evidences re-enforce the evidence accumulated from within the hyperlinked and hierarchical XML.
Scalability – Extrapolations

Constructing new data points outside a discrete set of known data points. Accelerating the convergence of the power method.

Aitken $\Delta^2$ Extrapolation

$$\tilde{x}^{(k-2)} = \tilde{u}_1 + \alpha_2 \tilde{u}_2$$

Quadratic Extrapolation

$$\tilde{x}^{(k-3)} = \tilde{u}_1 + \alpha_2 \tilde{u}_2 + \alpha_3 \tilde{u}_3$$

Power $A^d$

$$\lambda_2 = c$$

$$\tilde{x}^{(k-2)} = \tilde{u}_1 + \alpha_2 \tilde{u}_2 + \alpha_3 \tilde{u}_3$$
Scalability – Extrapolations

Constructing new data points outside a discrete set of known data points. Accelerating the convergence of the power method.

**Aitken \( \Delta^2 \) Extrapolation**

\[
\bar{x}^{(k-2)} = \vec{u}_1 + \alpha_2 \vec{u}_2
\]

**Quadratic Extrapolation**

\[
\bar{x}^{(k-3)} = \vec{u}_1 + \alpha_2 \vec{u}_2 + \alpha_3 \vec{u}_3
\]

**Power \( A^d \)**

\[
\lambda_2 = c
\]

\[
\bar{x}^{(k-2)} = \vec{u}_1 + \alpha_2 \vec{u}_2 + \alpha_3 \vec{u}_3
\]
Scalability – Extrapolations

Constructing new data points outside a discrete set of known data points. Accelerating the convergence of the power method.

**Aitken Δ² Extrapolation**

\[ \bar{x}^{(k-2)} = \bar{u}_1 + \alpha_2 \bar{u}_2 \]

**Quadratic Extrapolation**

\[ \bar{x}^{(k-3)} = \bar{u}_1 + \alpha_2 \bar{u}_2 + \alpha_3 \bar{u}_3 \]

**Power A^d**

\[ \lambda^2 = c \]

\[ \bar{x}^{(k-2)} = \bar{u}_1 + \alpha_2 \bar{u}_2 + \alpha_3 \bar{u}_3 \]
Scalability – Extrapolations

Constructing new data points outside a discrete set of known data points. Accelerating the convergence of the power method.

**Aitken Δ^2 Extrapolation**

\[ x^{(k-2)} = u_1 + \alpha_2 u_2 \]

**Quadratic Extrapolation**

\[ x^{(k-3)} = u_1 + \alpha_2 u_2 + \alpha_3 u_3 \]

**Power A^d**

\[ \lambda_2 = c \]

\[ x^{(k-2)} = u_1 + \alpha_2 u_2 + \alpha_3 u_3 \]
Selection Fusion

- Semi-structured retrieval is a combination of ranking and selection.
- Two fundamental simultaneous tasks.
- Ranking in itself not sufficient for the end users.
- Ranking and selection are distinct tasks.
- Ranking(scoring) system ranks “all” elements of the database (thorough list).
- Selection scheme: set of explicitly defined elements. Depends on use case (e.g. mobile device, result presentation, ad hoc etc).
- Selection fusion ranks explicitly given set of elements.
- Measuring performance is trade-off between ranking and selection.
Introduction

Background

Research Results

Experimental Evaluations

Conclusions

Selection Fusion

- Semi-structured retrieval is a combination of ranking and selection.
- Two fundamental simultaneous tasks.
  - Ranking in itself not sufficient for the end users.
  - Ranking and selection are distinct tasks.
  - Ranking(scoring) system ranks “all” elements of the database (thorough list).
  - Selection scheme: set of explicitly defined elements. Depends on use case (e.g. mobile device, result presentation, ad hoc etc).
  - Selection fusion ranks explicitly given set of elements.
  - Measuring performance is trade-off between ranking and selection.
Selection Fusion

- Semi-structured retrieval is a combination of *ranking* and *selection*.
- Two fundamental simultaneous tasks.
- Ranking in itself not sufficient for the end users.
  - Ranking and selection are distinct tasks.
  - Ranking (scoring) system ranks “all” elements of the database (thorough list).
  - Selection scheme: set of explicitly defined elements. Depends on use case (e.g. mobile device, result presentation, ad hoc etc).
  - Selection fusion ranks explicitly given set of elements.
  - Measuring performance is trade-off between ranking and selection.
Selection Fusion

- Semi-structured retrieval is a combination of ranking and selection.
- Two fundamental simultaneous tasks.
- Ranking in itself not sufficient for the end users.
- Ranking and selection are distinct tasks.
  - Ranking (scoring) system ranks “all” elements of the database (thorough list).
  - Selection scheme: set of explicitly defined elements. Depends on use case (e.g. mobile device, result presentation, ad hoc etc).
  - Selection fusion ranks explicitly given set of elements.
- Measuring performance is trade-off between ranking and selection.
Selection Fusion

- Semi-structured retrieval is a combination of ranking and selection.
- Two fundamental simultaneous tasks.
- Ranking in itself not sufficient for the end users.
- Ranking and selection are distinct tasks.
- Ranking (scoring) system ranks “all” elements of the database (thorough list).
  - Selection scheme: set of explicitly defined elements. Depends on use case (e.g. mobile device, result presentation, ad hoc etc).
- Selection fusion ranks explicitly given set of elements.
- Measuring performance is trade-off between ranking and selection.
Selection Fusion

- Semi-structured retrieval is a combination of ranking and selection.
- Two fundamental simultaneous tasks.
- Ranking in itself not sufficient for the end users.
- Ranking and selection are distinct tasks.
- Ranking (scoring) system ranks “all” elements of the database (thorough list).
- Selection scheme: set of explicitly defined elements. Depends on use case (e.g. mobile device, result presentation, ad hoc etc).
- Selection fusion ranks explicitly given set of elements.
- Measuring performance is trade-off between ranking and selection.
Selection Fusion

- Semi-structured retrieval is a combination of ranking and selection.
- Two fundamental simultaneous tasks.
- Ranking in itself not sufficient for the end users.
- Ranking and selection are distinct tasks.
- Ranking (scoring) system ranks “all” elements of the database (thorough list).
- Selection scheme: set of explicitly defined elements. Depends on use case (e.g. mobile device, result presentation, ad hoc etc).
- Selection fusion ranks explicitly given set of elements.
- Measuring performance is trade-off between ranking and selection.
Selection Fusion

- Semi-structured retrieval is a combination of *ranking* and *selection*.
- Two fundamental simultaneous tasks.
- Ranking in itself not sufficient for the end users.
- Ranking and selection are distinct tasks.
- Ranking(scoring) system ranks “all” elements of the database (thorough list).
- Selection scheme: set of explicitly defined elements. Depends on use case (e.g. mobile device, result presentation, ad hoc etc).
- Selection fusion ranks explicitly given set of elements.
- Measuring performance is trade-off between ranking and selection.
Selection Fusion Example

1  <A>
2  Text A1
3  <B> Text B
4  <D> Text D </D>
5  <E> Text E </E>
6  </B>
7  Text A2
8  <C> Text C
9  <F> Text F </F>
10 <G> Text G </G>
11 </C>
12 Text A3
13 </A>
Selection Fusion Example

The Contextual Features in Schema-Agnostic Environment
Selection Fusion Example

Ranking

\(< C,A,G,F,B,D,E >\)
Selection Fusion Example

Ranking
\(<C,A,G,F,B,D,E>\)

Selection Scheme
\(\{B,C\}\)
Selection Fusion Example

Ranking
\(<C, A, G, F, B, D, E>\)

Selection Scheme
\(\{B, C\}\)

Final Results
\(<C, B>\)
Outline

1. Introduction and Motivations
2. Background and Challenges
3. Research Results
4. Experimental Evaluations
5. Conclusions
Contributions

C1: Theoretical background of the area.
Contributions

C2: Extrapolation techniques and the inclusion of LAR based techniques in semi-structured retrieval. Although this latter proposition was not clearly specified in any of the papers, it is indirectly used in most of the work.
Contributions

**C4:** Contextualization using the structural context on the bibliographical structure of documents together with the Markovian random walk principle.
Contributions

C5: Contextualization of the hierarchical and hyperlink structures of documents in granulation selection and evaluation scenarios.
Contributions

**C6:** Worst-case analyses of the use of the structural context and the situation under which structural context might be beneficial.
Contributions

C7: Selection fusion—a methodology leading to a general semi-structured retrieval system, capable of serving a wide range of use case scenarios.
## Research Questions × Contributions

<table>
<thead>
<tr>
<th>Research questions</th>
<th>Papers</th>
<th>Contributions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>C1</td>
<td>C2</td>
</tr>
<tr>
<td>RQ</td>
<td>A,B</td>
<td></td>
</tr>
<tr>
<td>RQ1</td>
<td>D,F,G</td>
<td></td>
</tr>
<tr>
<td>RQ2</td>
<td>C,D,E,G</td>
<td></td>
</tr>
<tr>
<td>RQ3</td>
<td>D,E,G</td>
<td></td>
</tr>
<tr>
<td>RQ4</td>
<td>G</td>
<td></td>
</tr>
<tr>
<td>RQ_{ef}</td>
<td>B,C,D,E,F,G</td>
<td></td>
</tr>
</tbody>
</table>
Outline

1. Introduction and Motivations
2. Background and Challenges
3. Research Results
4. Experimental Evaluations
5. Conclusions
Datasets

Semantically annotated Wikipedia Collection

- 68 queries.
- 2,666,190 articles in XML.
- \( \approx 135 \) million hyperlinks.
- All the runs from INEX 2009 and 2010 submissions.
Datasets

Semantically annotated Wikipedia Collection

- 68 queries.
- 2,666,190 articles in XML.
- \( \approx 135 \) million hyperlinks.
- All the runs from INEX 2009 and 2010 submissions.

iSearch test Collection

Comprises scientific documents from the domain of physics.
\( \approx 500,000 \) papers.
- 18,443 book records in XML.
- 291,246 metadata for articles.
- 143,571 full-text articles in PDF.
- \( \approx 3.8 \) million bibliographical citations among the collection.
Baseline systems

- INDRI
- Different data fusion techniques applied on the runs submitted at INEX. Reciprocal rank.
- CombSUM\_Reciprocal.
- Individual INEX runs.
A representative set of baseline systems – task wise

<table>
<thead>
<tr>
<th>Run ID</th>
<th>Task</th>
<th>Papers</th>
<th>MAiP</th>
<th>MAgP</th>
<th>MAP</th>
<th>iP[0.01]</th>
<th>P5</th>
<th>rPrec</th>
</tr>
</thead>
<tbody>
<tr>
<td>indri</td>
<td>ad hoc</td>
<td>C</td>
<td>–</td>
<td>–</td>
<td>.0803</td>
<td>–</td>
<td>.1938</td>
<td>.1041</td>
</tr>
<tr>
<td>Reciprocal_Fusion</td>
<td>gran_article</td>
<td>D</td>
<td>–</td>
<td>–</td>
<td>.4141</td>
<td>–</td>
<td>.6618</td>
<td>–</td>
</tr>
<tr>
<td>UAmST</td>
<td>gran_article</td>
<td>D</td>
<td>–</td>
<td>–</td>
<td>.3578</td>
<td>–</td>
<td>.6500</td>
<td>–</td>
</tr>
<tr>
<td>Reciprocal_Fusion</td>
<td>gran_para</td>
<td>D</td>
<td>–</td>
<td>–</td>
<td>.2189</td>
<td>–</td>
<td>.4500</td>
<td>.3479</td>
</tr>
<tr>
<td>QTau</td>
<td>gran_para</td>
<td>D</td>
<td>–</td>
<td>–</td>
<td>.2286</td>
<td>–</td>
<td>.5324</td>
<td>.2779</td>
</tr>
<tr>
<td>CombSUM_Reciprocal</td>
<td>thor-foc</td>
<td>EG</td>
<td>.3396</td>
<td>–</td>
<td>–</td>
<td>.7273</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>UWFeRBM25F2</td>
<td>thor-foc</td>
<td>EG</td>
<td>.1854</td>
<td>–</td>
<td>–</td>
<td>.6333</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>I09LIP6Okapi</td>
<td>thor-foc</td>
<td>EG</td>
<td>.3000</td>
<td>–</td>
<td>–</td>
<td>.6141</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>UJM_15525</td>
<td>thor-foc</td>
<td>EG</td>
<td>.2890</td>
<td>–</td>
<td>–</td>
<td>.6060</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>UamsFSecs2dbi100CA</td>
<td>thor-foc</td>
<td>EG</td>
<td>.1928</td>
<td>–</td>
<td>–</td>
<td>.5997</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>BM25BOTrangeFOC</td>
<td>thor-foc</td>
<td>EG</td>
<td>.2912</td>
<td>–</td>
<td>–</td>
<td>.5992</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Spirix09R001</td>
<td>thor-foc</td>
<td>EG</td>
<td>.2865</td>
<td>–</td>
<td>–</td>
<td>.5903</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>LIG-2009-focused-1F</td>
<td>thor-foc</td>
<td>EG</td>
<td>.2702</td>
<td>–</td>
<td>–</td>
<td>.5853</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>BM25AncestorBIC</td>
<td>bic</td>
<td>G</td>
<td>–</td>
<td>–</td>
<td>.1706</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>BM25AncestorRIC</td>
<td>ric</td>
<td>G</td>
<td>–</td>
<td>–</td>
<td>.1865</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>UamsRSCMACMdbi100</td>
<td>ric</td>
<td>G</td>
<td>–</td>
<td>–</td>
<td>.1773</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>UamsFSsec2dbi100CA</td>
<td>foc</td>
<td>G</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>.5997</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>UAmSL09article</td>
<td>thor</td>
<td>G</td>
<td>.2818</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>
Outline

1 Introduction and Motivations
2 Background and Challenges
3 Research Results
4 Experimental Evaluations
5 Conclusions
Take away

2. Extrapolation techniques.
3. Unveil the role of the structural context.
4. Utilization of the structural context not ‘in isolation’ but ‘in contextualization’.
5. Contextualization with random walk as a theoretically sound model.
6. Selection fusion: a fusion of two distinct and independent systems.
7. Developed competitive baseline systems.
8. Evaluated them using state of the art benchmarks.
Impact

1. How semantics in the document, i.e., the structure, could possibly be used?
2. How do the evidences originate and propagate across the structure?
3. Use of context from internal and external contextual features to improve retrieval.
4. Ranking and Selection could be done as an independent task.
5. Good ranking is successful in various selection comparisons.
6. Most of the proposed systems are empirically evaluated and found to be statistically sound.
Thank you for your attention!