CIIR CENTER FOR INTELLIGENT INFORMATION RETRIEVAL
UNIVERSITY OF MASSACHUSETTS AMHERST

Parameterized Concept Weighting for Information Retrieval

Michael Bendersky

Joint Work with

W. Bruce Croft, *UMass Amherst*

Donald Metzler, *ISI USC*

David A. Smith, *UMass Amherst*

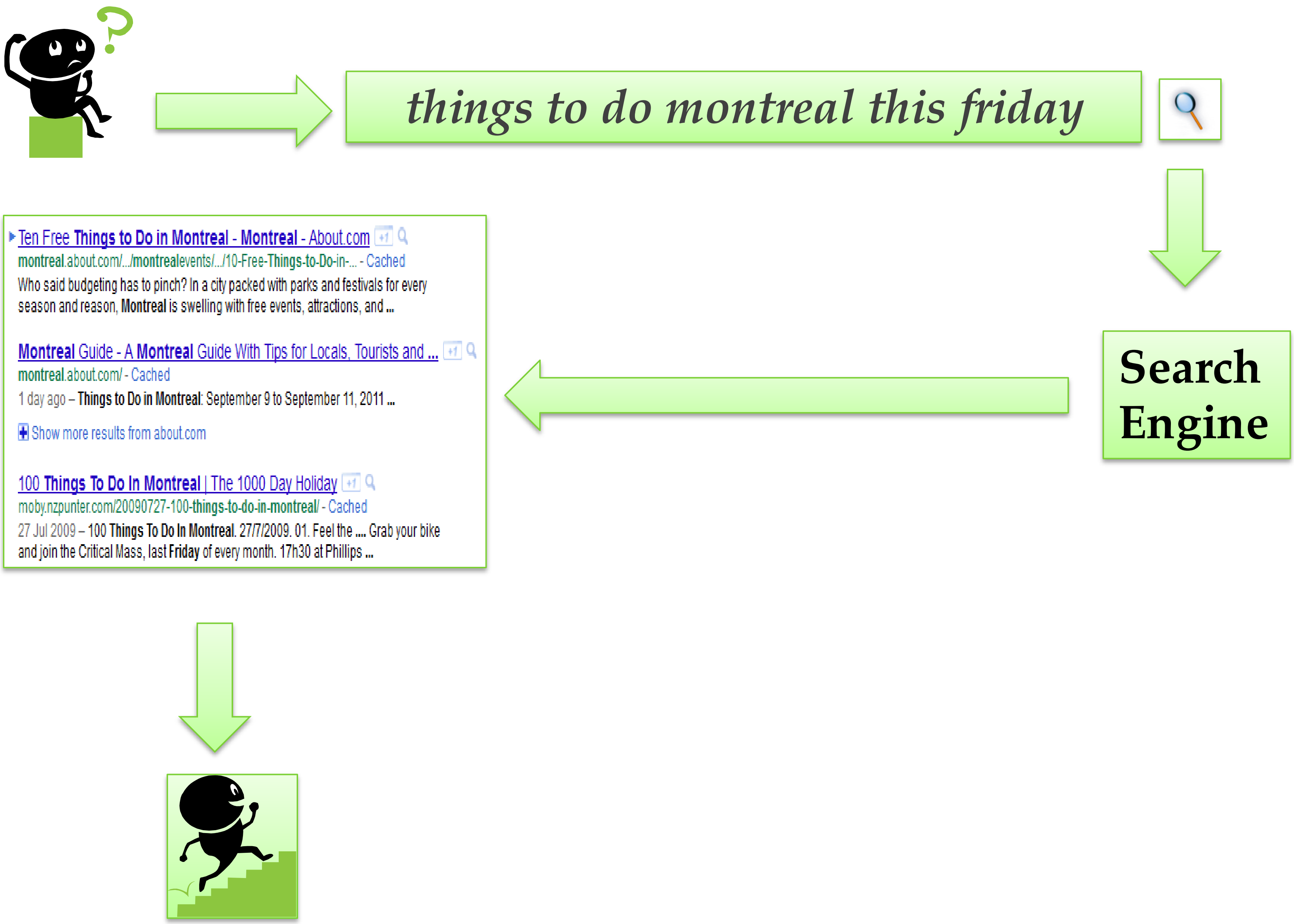
Université de Montréal, Sept. 2011

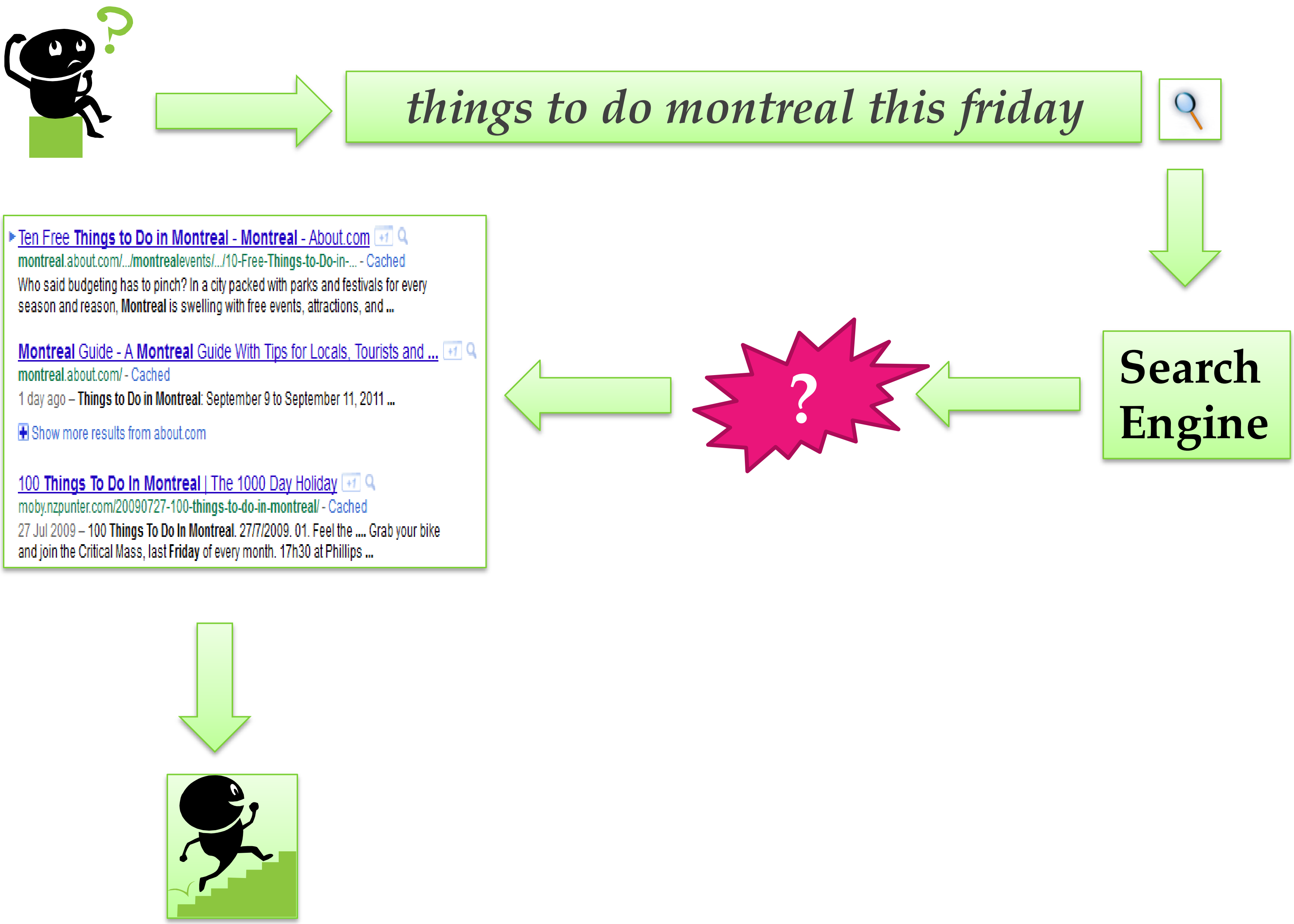
Talk Outline

1. Search Query Representation
2. Parameterized Concept Weighting
3. Explicit Concept Weighting
4. Expansion Concept Weighting
5. Concept Weighting on Web Scale

1. **Search Query Representation**
2. Parameterized Concept Weighting
3. Explicit Concept Weighting
4. Expansion Concept Weighting
5. Concept Weighting on Web Scale

SEARCH QUERY REPRESENTATION


things to do montreal this friday
Search Engine
Ten Free Things to Do in Montreal - Montreal - About.com
montreal.about.com/.../montrealevents/.../10-Free-Things-to-Do-in-... - Cached
Who said budgeting has to pinch? In a city packed with parks and festivals for every season and reason, Montreal is swelling with free events, attractions, and ...
Montreal Guide - A Montreal Guide With Tips for Locals, Tourists and ...
montreal.about.com/ - Cached
1 day ago – Things to Do in Montreal: September 9 to September 11, 2011 ...
Show more results from about.com
100 Things To Do In Montreal | The 1000 Day Holiday
moby.nzpunter.com/20090727-100-things-to-do-in-montreal/ - Cached
27 Jul 2009 – 100 Things To Do In Montreal. 27/7/2009. 01. Feel the .... Grab your bike and join the Critical Mass, last Friday of every month. 17h30 at Phillips ...



things to do montreal this friday
Search Engine
?
Ten Free Things to Do in Montreal - Montreal - About.com
montreal.about.com/.../montrealevents/.../10-Free-Things-to-Do-in-... - Cached
Who said budgeting has to pinch? In a city packed with parks and festivals for every season and reason, Montreal is swelling with free events, attractions, and ...
Montreal Guide - A Montreal Guide With Tips for Locals, Tourists and ...
montreal.about.com/ - Cached
1 day ago – Things to Do in Montreal: September 9 to September 11, 2011 ...
Show more results from about.com
100 Things To Do In Montreal | The 1000 Day Holiday
moby.nzpunter.com/20090727-100-things-to-do-in-montreal/ - Cached
27 Jul 2009 – 100 Things To Do In Montreal. 27/7/2009. 01. Feel the .... Grab your bike and join the Critical Mass, last Friday of every month. 17h30 at Phillips ...

The Challenges of Query Representation

things to do montreal this friday

- **The linguistic structure of the query is never explicitly observed**
- Structure inference is hard
  - *Short and ambiguous search query*
  - *Idiosyncratic grammar*
  - *No capitalization and punctuation*
- Strict limit on inference time

The Challenges of Query Representation

things to do montreal this friday

- The linguistic structure of the query is never explicitly observed
- **Structure inference is hard**
  - *Short and ambiguous search query*
  - *Idiosyncratic grammar*
  - *No capitalization and punctuation*
- Strict limit on inference time

The Challenges of Query Representation

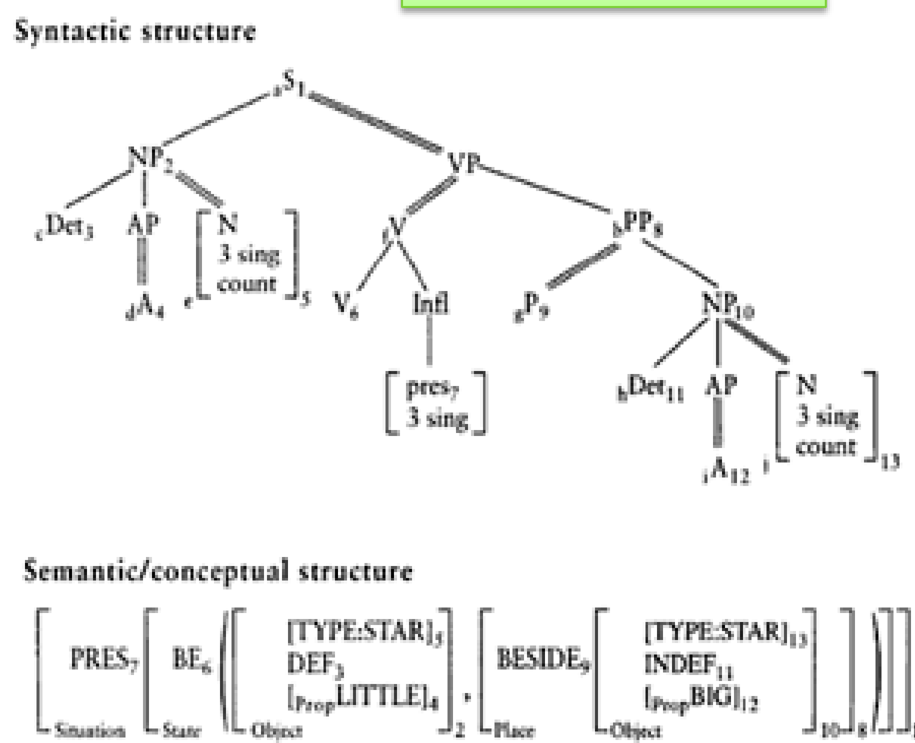
things to do montreal this friday

- The linguistic structure of the query is never explicitly observed
- Structure inference is hard
  - *Short and ambiguous search query*
  - *Idiosyncratic grammar*
  - *No capitalization and punctuation*
- **Strict limit on inference time**

Query Representations Spectrum

Too coarse

Too fine grained



Applications of Complex Query Representations

- **Verbose queries in web search**
  (Experian Hitwise report, 2010)
  - *Growth of **5+** word queries since 2008 – **15%***
  - *Total share of the query traffic – **20%***
- Emerging search modalities
  - *Voice – activated search*
  - *Search on mobile devices*
- Q&A systems
- Enterprise & Academic Search

Applications of Complex Query Representations

- Verbose queries in web search
  (Experian Hitwise report, 2010)
  - *Growth of **5+** word queries since 2008 – **15%***
  - *Total share of the query traffic – **20%***
- **Emerging search modalities**
  - *Voice – activated search*
  - *Search on mobile devices*
- Q&A systems
- Enterprise & Academic Search

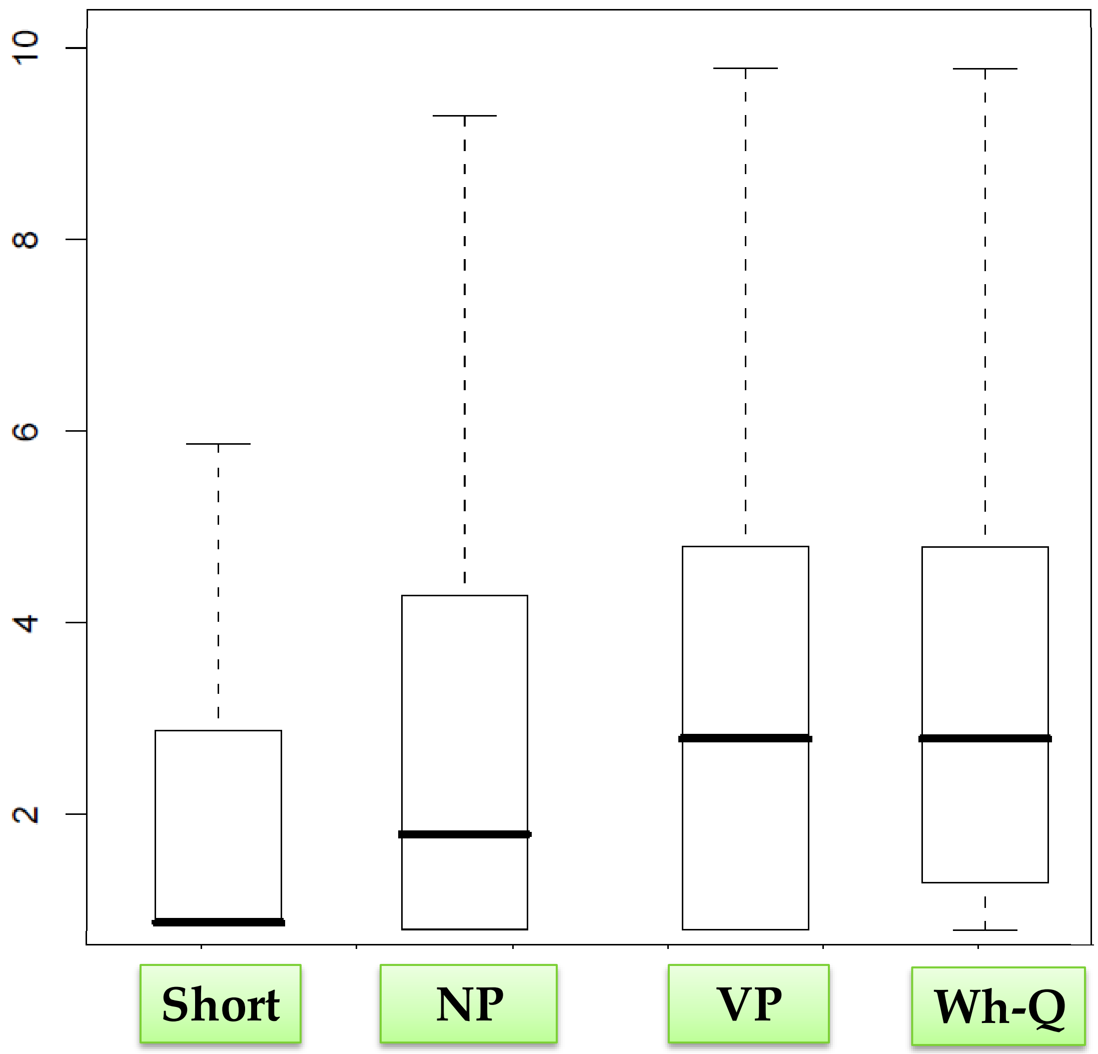
Applications of Complex Query Representations

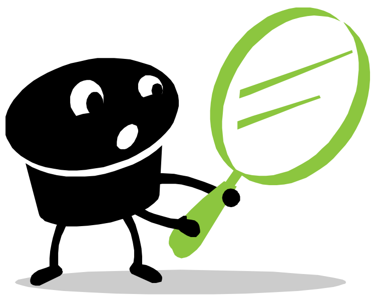
- Verbose queries in web search
  (Experian Hitwise report, 2010)
  - *Growth of **5+** word queries since 2008 – **15%***
  - *Total share of the query traffic – **20%***
- Emerging search modalities
  - *Voice – activated search*
  - *Search on mobile devices*
- **Q&A systems**
- Enterprise & Academic Search

Applications of Complex Query Representations

- Verbose queries in web search
  (Experian Hitwise report, 2010)
  - *Growth of **5+** word queries since 2008 – **15%***
  - *Total share of the query traffic – **20%***
- Emerging search modalities
  - *Voice – activated search*
  - *Search on mobile devices*
- Q&A systems
- **Enterprise & Academic Search**

Query Difficulty by Type
(Bendersky & Croft, 2009)

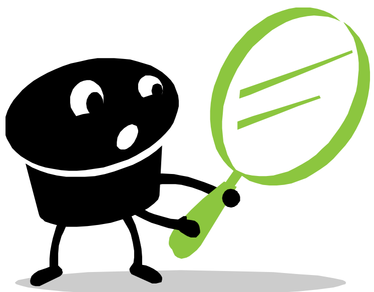




volcano eruptions effect global temperature

Words

volcano
eruptions
effect
global
temperature



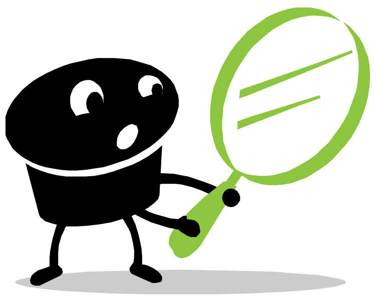
volcano eruptions effect global temperature

Words

volcano
eruptions
effect
global
temperature

Phrases

volcano eruptions
global temperature



volcano eruptions effect global temperature

Words

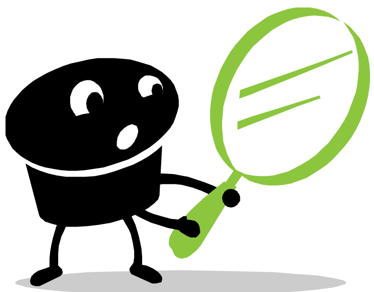
volcano
eruptions
effect
global
temperature

Phrases

volcano eruptions
global temperature

Expansion

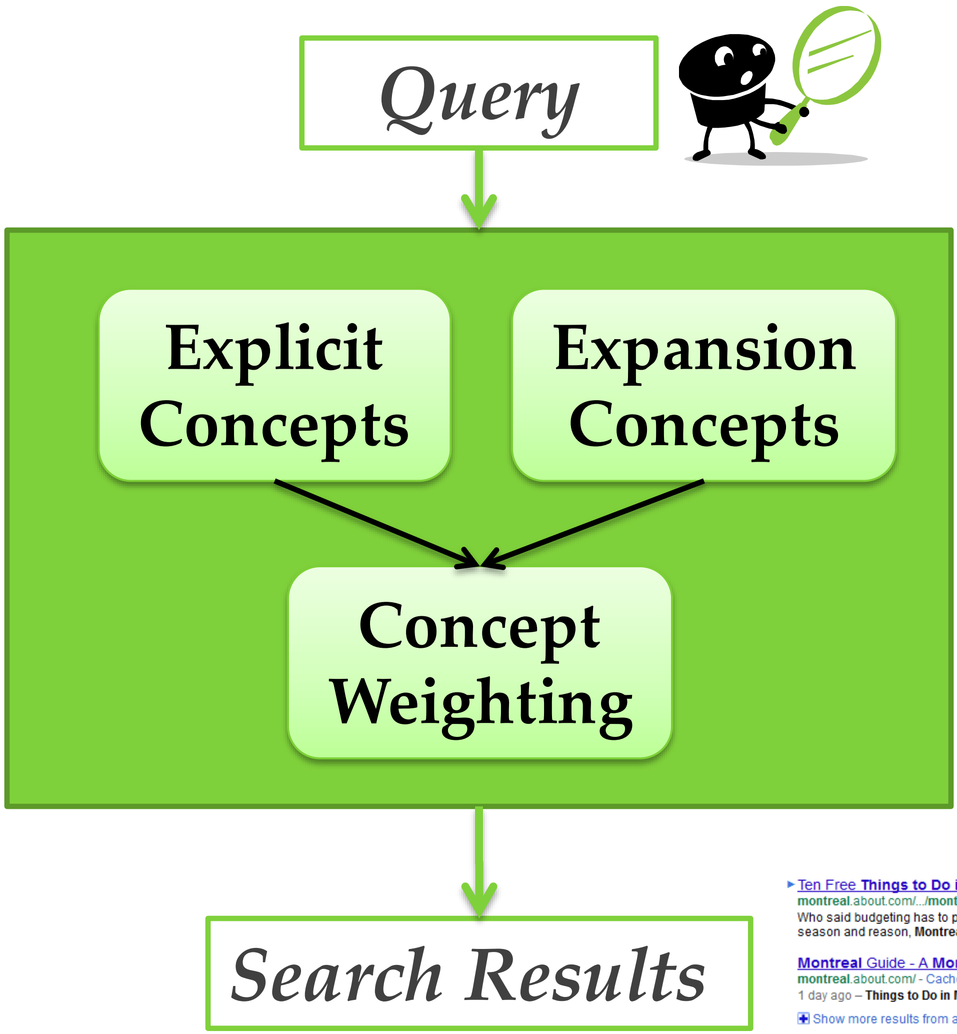
ash
climate
earth
lava
…

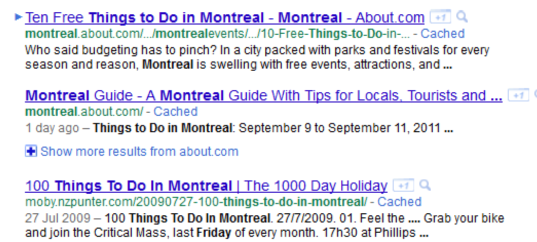


volcano eruptions effect global temperature

Words	Phrases	Expansion
volcano	volcano eruptions	ash
eruptions	global temperature	climate
effect		earth
global		lava
temperature		…

Query Representation Process





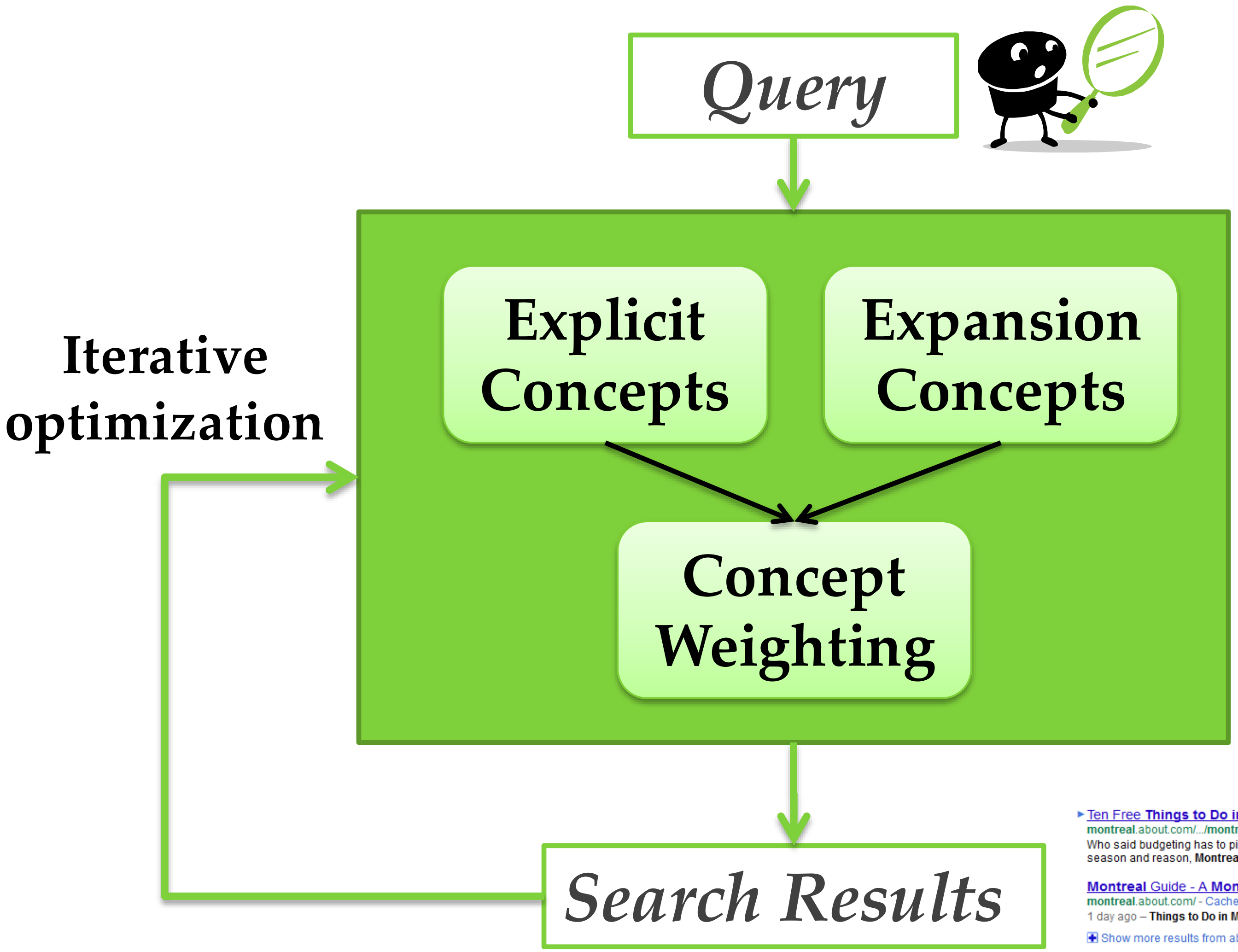
Ten Free Things to Do in Montreal - Montreal - About.com
montreal.about.com/.../montrealevents/.../10-Free-Things-to-Do-in-... - Cached
Who said budgeting has to pinch? In a city packed with parks and festivals for every season and reason, Montreal is swelling with free events, attractions, and ...

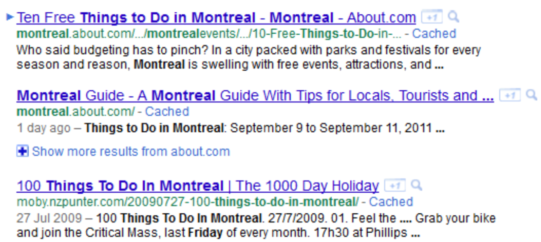
Montreal Guide - A Montreal Guide With Tips for Locals, Tourists and ...
montreal.about.com/ - Cached
1 day ago – Things to Do in Montreal: September 9 to September 11, 2011 ...

Show more results from about.com

100 Things To Do In Montreal | The 1000 Day Holiday
moby.nzpunter.com/20090727-100-things-to-do-in-montreal/ - Cached
27 Jul 2009 – 100 Things To Do In Montreal. 27/7/2009. 01. Feel the .... Grab your bike and join the Critical Mass, last Friday of every month. 17h30 at Phillips ...

Query Representation Process





Query Representations in IR: *Unsupervised Term Weighting*

- The majority of common **bag-of-words** models use unsupervised term weighting
  - **BM25** *(Robertson & Walker 1994)*
  - **Query Likelihood** *(Ponte & Croft 1998)*
  - **Divergence from Randomness** *(Amati & Van Rijsbergen 2002)*

- **Inverse Document Frequency (*IDF*)** is a popular term weighting measure

$$IDF(t) = \log \frac{|D|}{|\{d : t \in d\}|}$$

Query Representations in IR: *Supervised Term Weighting*

- More recent work explores the importance of supervised term weighting
  - Going beyond *IDF*
- Focus on verbose queries
  - **Regression Rank** *(Lease 2009)*
  - **Term Selection** *(Lee et al. 2009)*
  - **Term Necessity** *(Zhao & Callan 2010)*

Query Representations in IR: *Supervised Concept Weighting*

- Focus on a specific concept type
  - **Noun Phrases**
    (Bendersky & Croft 2008)
  - **Phrases & Proximities**
    (Bendersky & Croft 2010, Shi & Nie 2010)
  - **Term Spans**
    (Svore et al. 2010)

Query Representations in IR: *Supervised Expansion Weighting*

- Most common query expansion approaches use unsupervised weighting
- *Cao et al. (2008)* use binary classification for expansion term weighting
  - No supervised weighting for explicit query concepts

PARAMETERIZED CONCEPT WEIGHTING

Concepts – Semantic Definition

An abstract idea or a mental symbol defined as a "unit of knowledge"

- **General, non-operational definition**
- **Should be adapted based on the application domain**

Concepts – Information Retrieval Definition

Any syntactic expression that can be matched within a document

- **A broad definition that is able to capture a variety of linguistic phenomena**
- **Easy to use in retrieval models**
- **Practical generalization of the semantic definition**

Concept Types

- $\mathcal{T}$ – set of possible concept types
  - *Query terms*
  - *Exact phrases*
  - *Proximity matches*
  - *Expansion terms from the corpus*
  - *Expansion terms from external sources*
  - …

In this talk

Concept-Based Retrieval

$$sc(Q, D) = \sum_{T \in \mathcal{T}} \sum_{\kappa \in T} \lambda_\kappa f(\kappa, D)$$

Concept Types

Concept-Based Retrieval

$$sc(Q, D) = \sum_{T \in \mathcal{T}} \sum_{\kappa \in T} \lambda_\kappa f(\kappa, D)$$

Concepts

Concept-Based Retrieval

$$sc(Q,D) = \sum_{T \in \mathcal{T}} \sum_{\kappa \in T} \lambda_\kappa \boxed{f(\kappa, D)}$$

Matching Function

$$f(\kappa, D) = \log \frac{tf_{\kappa,D} + \mu \frac{tf_{\kappa,\mathcal{C}}}{|\mathcal{C}|}}{|D| + \mu}$$

Language Modeling Estimate with Dirichlet Smoothing
(Zhai & Lafferty, 2001)

Concept-Based Retrieval

$$sc(Q, D) = \sum_{T \in \mathcal{T}} \sum_{\kappa \in T} \boxed{\lambda_\kappa} f(\kappa, D)$$

Concept Weight

Estimating Concept Weights

$$sc(Q, D) = \sum_{T \in \mathcal{T}} \sum_{\kappa \in T} \boxed{\lambda_\kappa} f(\kappa, D)$$

Option I

Tying the weights λ_κ for concepts of type **T**

All the concepts of the same type are equally important for expressing query intent

Estimating Concept Weights

$$sc(Q, D) = \sum_{T \in \mathcal{T}} \sum_{\kappa \in T} \boxed{\lambda_\kappa} f(\kappa, D)$$

Option II

Separately estimating λ_κ for each concept κ

Infeasible – the number of possible concepts is exponential in the size of the vocabulary.

Estimating Concept Weights

$$sc(Q, D) = \sum_{T \in \mathcal{T}} \sum_{\kappa \in T} \boxed{\lambda_\kappa} f(\kappa, D)$$

<u>**Option III**</u>

Parameterizing the weights λ_κ

Parameterize a concept of type T using a set of importance features Φ^T

Weight Parameterization

$$sc(Q,D) = \sum_{T \in \mathcal{T}} \sum_{\kappa \in T} \boxed{\lambda_\kappa} f(\kappa, D)$$

Parameterized Weight

$$\boxed{\lambda_\kappa = \sum_{\varphi \in \Phi^T} w_\varphi \varphi(\kappa)}$$

Weight Parameterization

$$sc(Q,D) = \sum_{T \in \mathcal{T}} \sum_{\kappa \in T} \lambda_\kappa f(\kappa, D)$$

Parameterized Weight

$$\boxed{\lambda_\kappa = \sum_{\varphi \in \Phi^T} w_\varphi \varphi(\kappa)}$$

Concept Importance Feature

Weight Parameterization

$$sc(Q, D) = \sum_{T \in \mathcal{T}} \sum_{\kappa \in T} \lambda_\kappa f(\kappa, D)$$

Parameterized Weight

$$\lambda_\kappa = \sum_{\varphi \in \Phi^T} w_\varphi \varphi(\kappa)$$

Feature Weight

Concept Importance Feature

Concept Importance Features

Feature	Description
GF(κ)	Frequency of concept κ in Google n-grams
WF(κ)	Frequency of concept κ in Wikipedia titles
QF(κ)	Frequency of concept κ in a search log
CF(κ)	Frequency of concept κ in the collection
DF(κ)	Document frequency of concept κ
AP(κ)	A priori concept weight

Parameterized Retrieval Model

$$sc(Q,D) = \sum_{T \in \mathcal{T}} \sum_{\varphi \in \Phi^{T}} w_{\varphi} \sum_{\kappa \in T} \varphi(\kappa) f(\kappa, D)$$

Concept Types

Parameterized Retrieval Model

$$sc(Q,D) = \sum_{T \in \mathcal{T}} \sum_{\varphi \in \Phi^{T}} w_{\varphi} \sum_{\kappa \in T} \varphi(\kappa) f(\kappa, D)$$

Concept Types

Importance Features

Parameterized Retrieval Model

$$sc(Q, D) = \sum_{T \in \mathcal{T}} \sum_{\varphi \in \Phi^{T}} w_{\varphi} \sum_{\kappa \in T} \varphi(\kappa) f(\kappa, D)$$

Concept Types

Importance Features

Concepts

Parameterized Retrieval Model

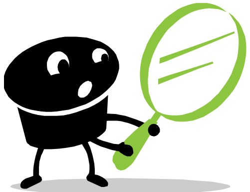
$$sc(Q,D) = \sum_{T \in \mathcal{T}} \sum_{\varphi \in \Phi^{T}} \boxed{w_{\varphi}} \sum_{\kappa \in T} \varphi(\kappa) f(\kappa, D)$$

- Linear in $\mathbf{W} = \{ \mathbf{w}_{\varphi} \mid \varphi \in \Phi^{T}, \mathrm{T} \in \mathcal{T} \}$
- Can be optimized using learning-to-rank techniques
  - **Coordinate Ascent** *(Metzler & Croft, 2007)*

EXPLICIT CONCEPT WEIGHTING

"Learning Concept Importance Using a Weighted Dependence Model"

(Bendersky et. al, WSDM 2010)



volcano eruptions effect global temperature

Words

volcano
eruptions
effect
global
temperature

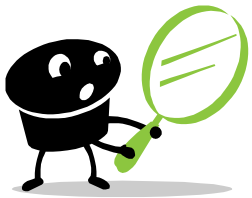
Exact Phrases

"volcano eruptions"
"eruptions effect"
"effect global"
"global temperature"

Proximity Matches

volcano...eruptions OR eruptions…volcano
eruptions…effect OR effect…eruptions
effect…global OR global…effect
global...temperature OR temperature…global

Sequential Dependence (*SD*) *(Metzler & Croft, 2007)*



volcano eruptions effect global temperature

Words

volcano
eruptions
effect
global
temperature

Exact Phrases

"volcano eruptions"
"eruptions effect"
"effect global"
"global temperature"

Proximity Matches

volcano...eruptions OR eruptions…volcano
eruptions…effect OR effect…eruptions
effect…global OR global…effect
global...temperature OR temperature…global

Weighted Sequential Dependence (*WSD*) *(Bendersky et. al, 2010)*

Learning Concept Weights in WSD

$$sc(Q,D) = \sum_{T \in \mathcal{T}} \sum_{\varphi \in \Phi^T} \boxed{w_\varphi} \sum_{\kappa \in T} \varphi(\kappa) f(\kappa, D)$$

- Initialize $\mathbf{W} = \{ \mathbf{w}_\varphi \mid \varphi \in \Phi^T, \mathrm{T} \in \mathcal{T} \}$
- While improvement in **MAP**
  - For each $\mathbf{w}_\varphi \in \mathbf{W}$
    - Line search for optimal value of $\mathbf{w}_\varphi$
    - At each search iteration test **MAP**

Learning Concept Weights in WSD

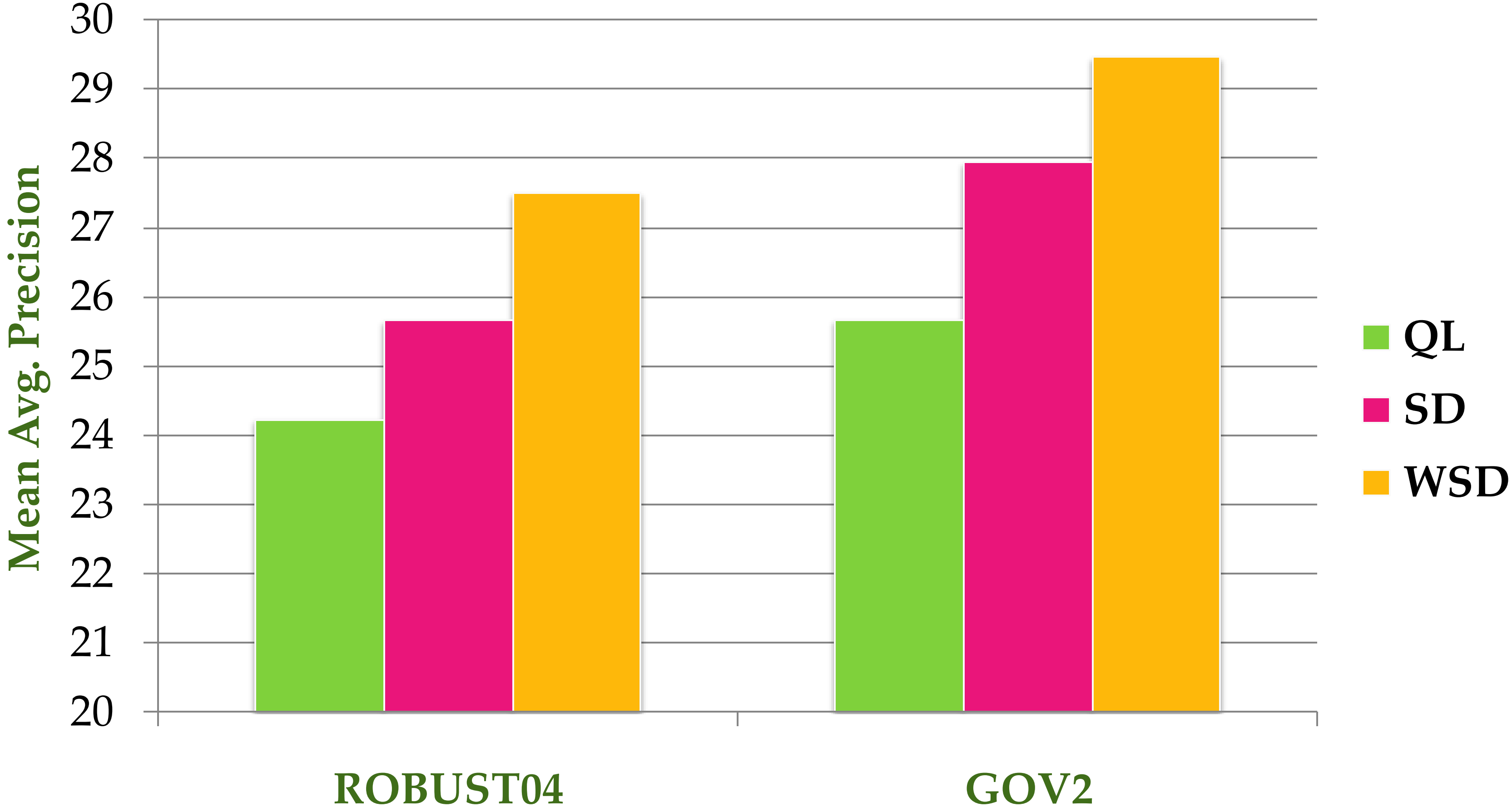
$$sc(Q,D) = \sum_{T \in \mathcal{T}} \sum_{\varphi \in \Phi^T} \boxed{w_\varphi} \sum_{\kappa \in T} \varphi(\kappa) f(\kappa, D)$$

- Initialize $\mathbf{W} = \{ w_\varphi \mid \varphi \in \Phi^T, T \in \mathcal{T} \}$
- While improvement in **MAP**
  - For each $\mathbf{w}_\varphi \in \mathbf{W}$
    - Line search for optimal value of $\mathbf{w}_\varphi$
    - At each search iteration test **MAP**

Comparison with Non-Parameterized Methods

	Query Terms	Exact Phrases	Proximity Matches
Query Likelihood (**QL**)	$\mathcal{N}$		
Sequential Dependence (**SD**)	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$
Weighted Sequential Dependence (**WSD**)	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$

Comparison with Non-Parameterized Methods



Newswire Collection

Crawl of .gov domain

1. Search Query Representation
2. Parameterized Concept Weighting
3. Explicit Concept Weighting
4. **Expansion Concept Weighting**
5. Concept Weighting on Web Scale

EXPANSION CONCEPT WEIGHTING

"Parameterized Concept Weighting in Verbose Queries"
(Bendersky et. al, SIGIR 2011)

Parameterized Query Expansion

$$sc(Q,D) = \sum_{\boxed{T \in \mathcal{T}}} \sum_{\varphi \in \Phi^T} w_\varphi \sum_{\kappa \in T} \varphi(\kappa) f(\kappa, D)$$

- **Explicit Query Concepts**
  - Terms
  - Exact Phrases
  - Proximity Matches
- **Expansion Terms**

Parameterized Query Expansion

$$sc(Q,D) = \sum_{T \in \mathcal{T}} \sum_{\varphi \in \Phi^T} \boxed{w_\varphi} \sum_{\kappa \in T} \varphi(\kappa) f(\kappa, D)$$

Linear in $\mathbf{W} = \{ \mathbf{w}_\varphi \mid \boldsymbol{\varphi} \in \boldsymbol{\Phi}^{\mathrm{T}}, \mathrm{T} \in \mathcal{T} \}$

Parameterized Query Expansion

$$sc(Q,D) = \underbrace{\sum_{\boxed{T \in \mathcal{T}}}}_{\text{Explicit Query Concepts, Expansion Terms}} \sum_{\varphi \in \Phi^T} \underbrace{\boxed{w_\varphi}}_{\text{Linear in } W = \{ w_\varphi \mid \varphi \in \Phi^T, T \in \mathcal{T} \}} \sum_{\kappa \in T} \varphi(\kappa) f(\kappa, D)$$

- Standard ranking optimization considers only **explicit query concepts**
- ***PQE*** combines evidence from both **explicit query concepts** and **expansion terms**
- Explicit concept weights impact the choice of expansion concepts

Latent Concept Expansion
(Metzler & Croft 2007)

"Camels in North America"

Explicit Concepts

weight	term
.8	camel
.8	north
.8	america
.2	"camel north"
.2	"north america"

Expansion with pseudo-relevance feedback

Expansion Terms

weight	term
.0178	indians
.0031	mexico
.0028	new
.0024	dress
.0021	clothing
…	…

Latent Concept Expansion
(Metzler & Croft 2007)

"Camels in North America"

Explicit Concepts

weight	term
.8	camel
.8	north
.8	america
.2	"camel north"
.2	"north america"

Expansion with pseudo-relevance feedback

Expansion Terms

weight	term
.0178	indians
.0031	mexico
.0028	new
.0024	dress
.0021	clothing
…	…

Mean Avg. Prec. 0.07

Parameterized Query Expansion

"Camels in North America"

Explicit Concepts

weight	term
.2591	camel
.1783	north
.1969	america
.0328	"camel north"
.0328	"north america"

Expansion with pseudo-relevance feedback

Expansion Terms

weight	term
.0314	bison
.0314	oil
.0306	nafta
.0305	fossil
.0269	expansion
…	…

Parameterized Query Expansion

"Camels in North America"

Explicit Concepts

weight	term
.2591	camel
.1783	north
.1969	america
.0328	"camel north"
.0328	"north america"

Expansion with pseudo-relevance feedback

Expansion Terms

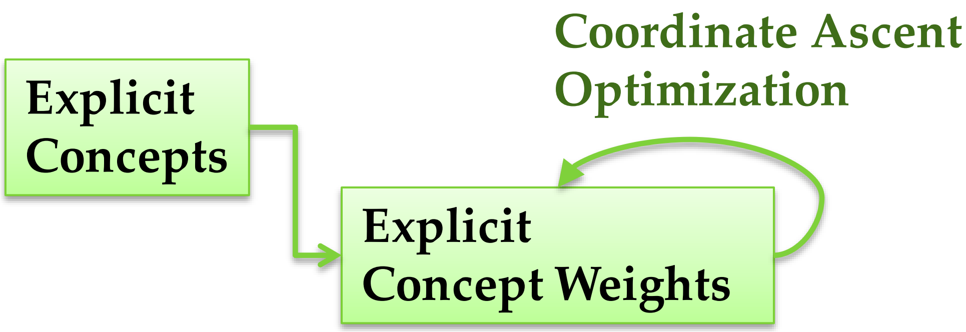
weight	term
.0314	bison
.0314	oil
.0306	nafta
.0305	fossil
.0269	expansion
…	…

Mean Avg. Prec. 0.49

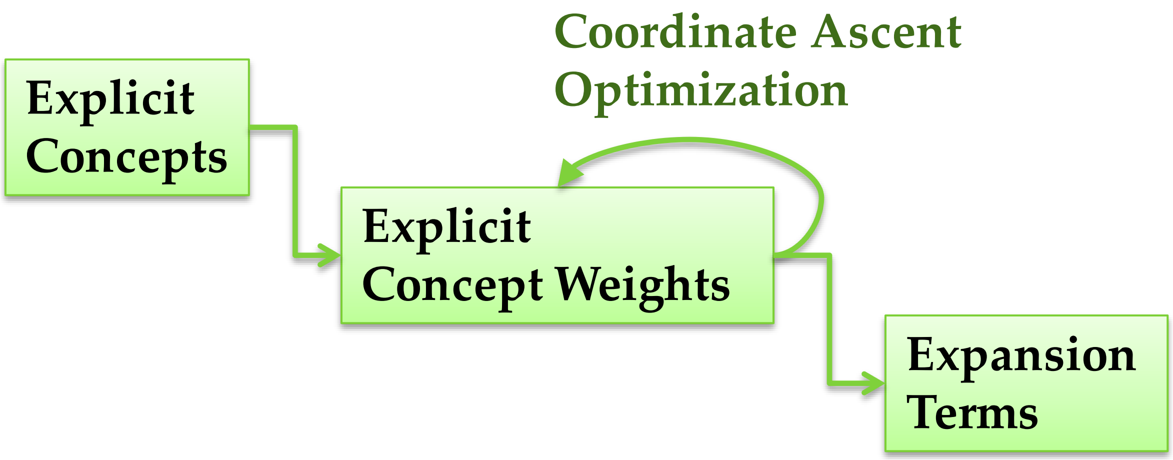
PQE Training Process

Explicit Concepts

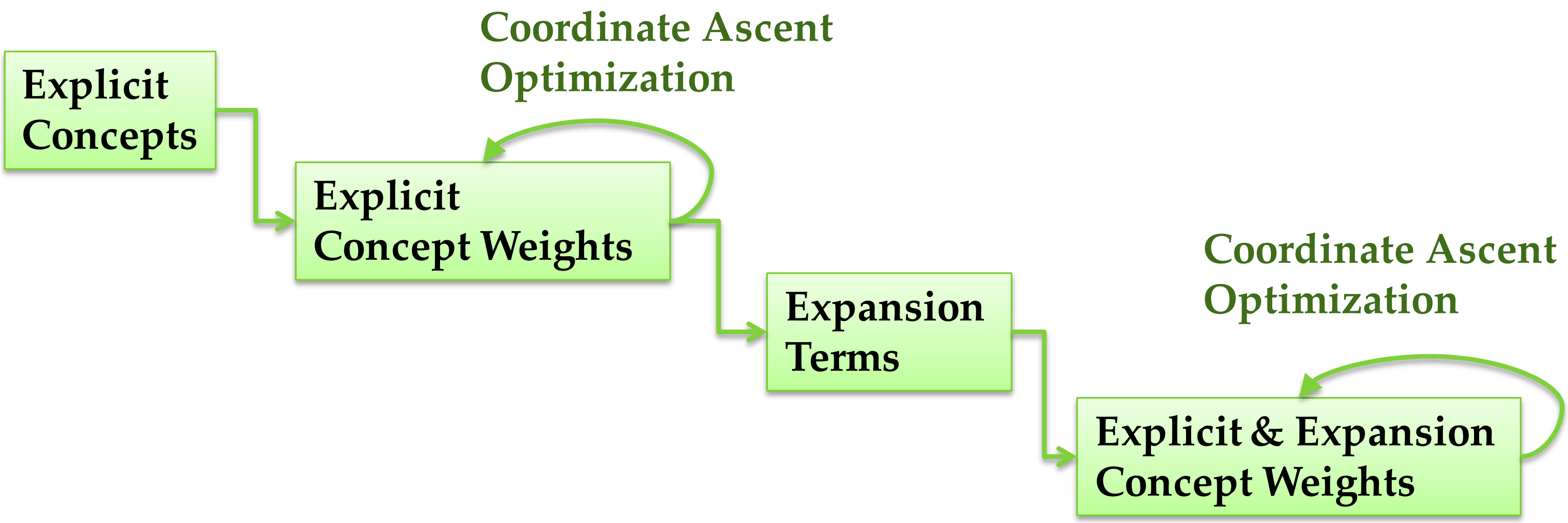
PQE Training Process



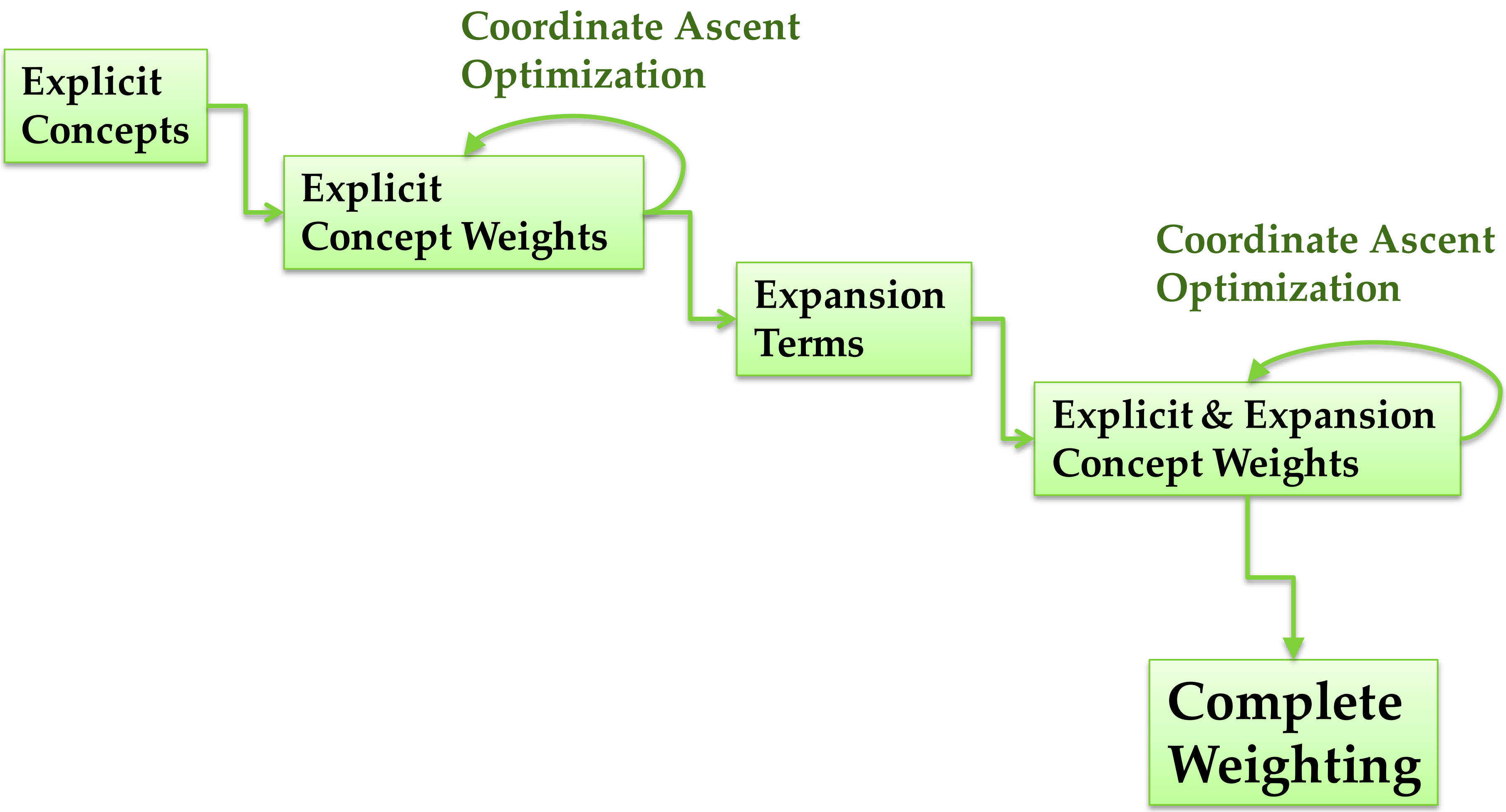
PQE Training Process



PQE Training Process



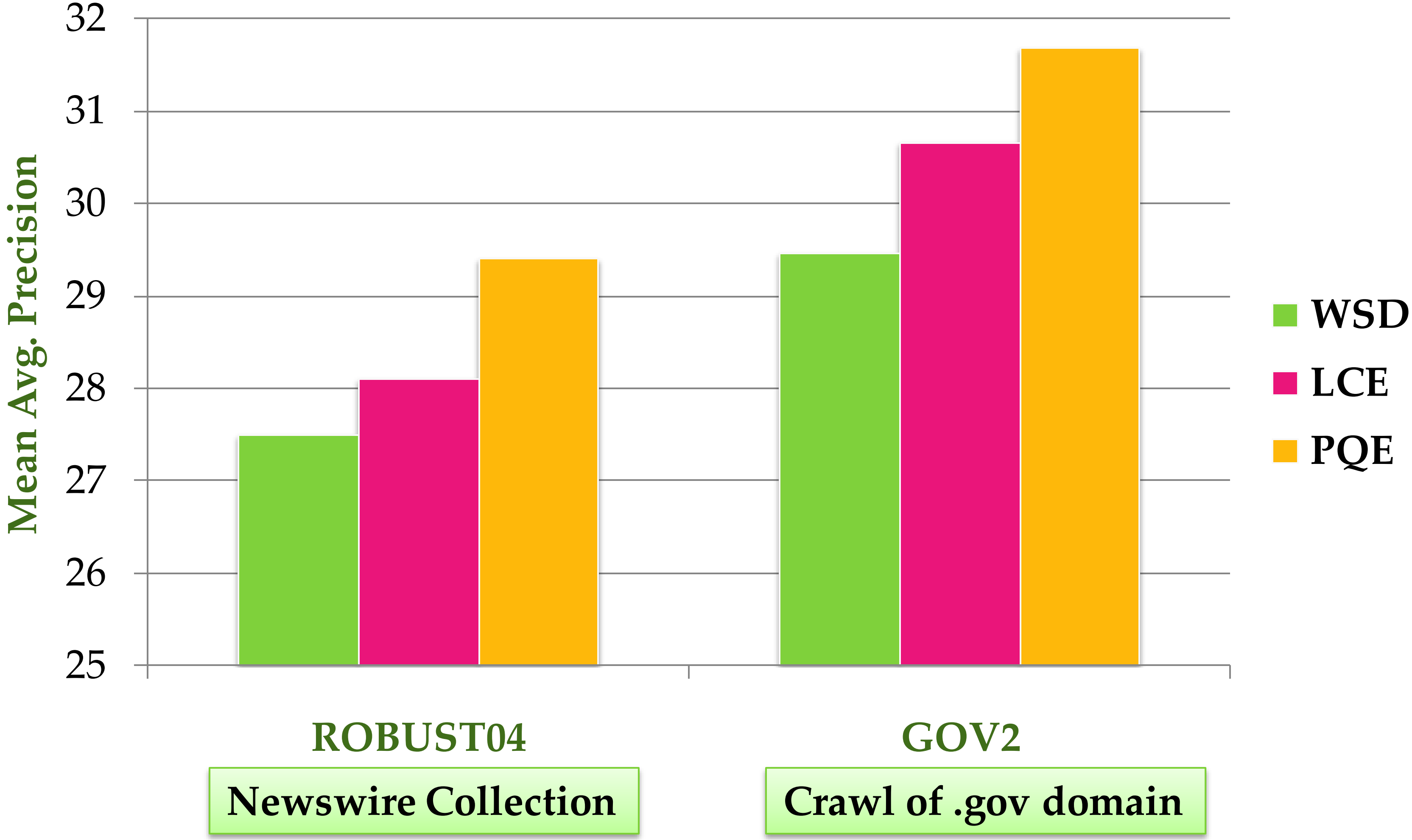
PQE Training Process



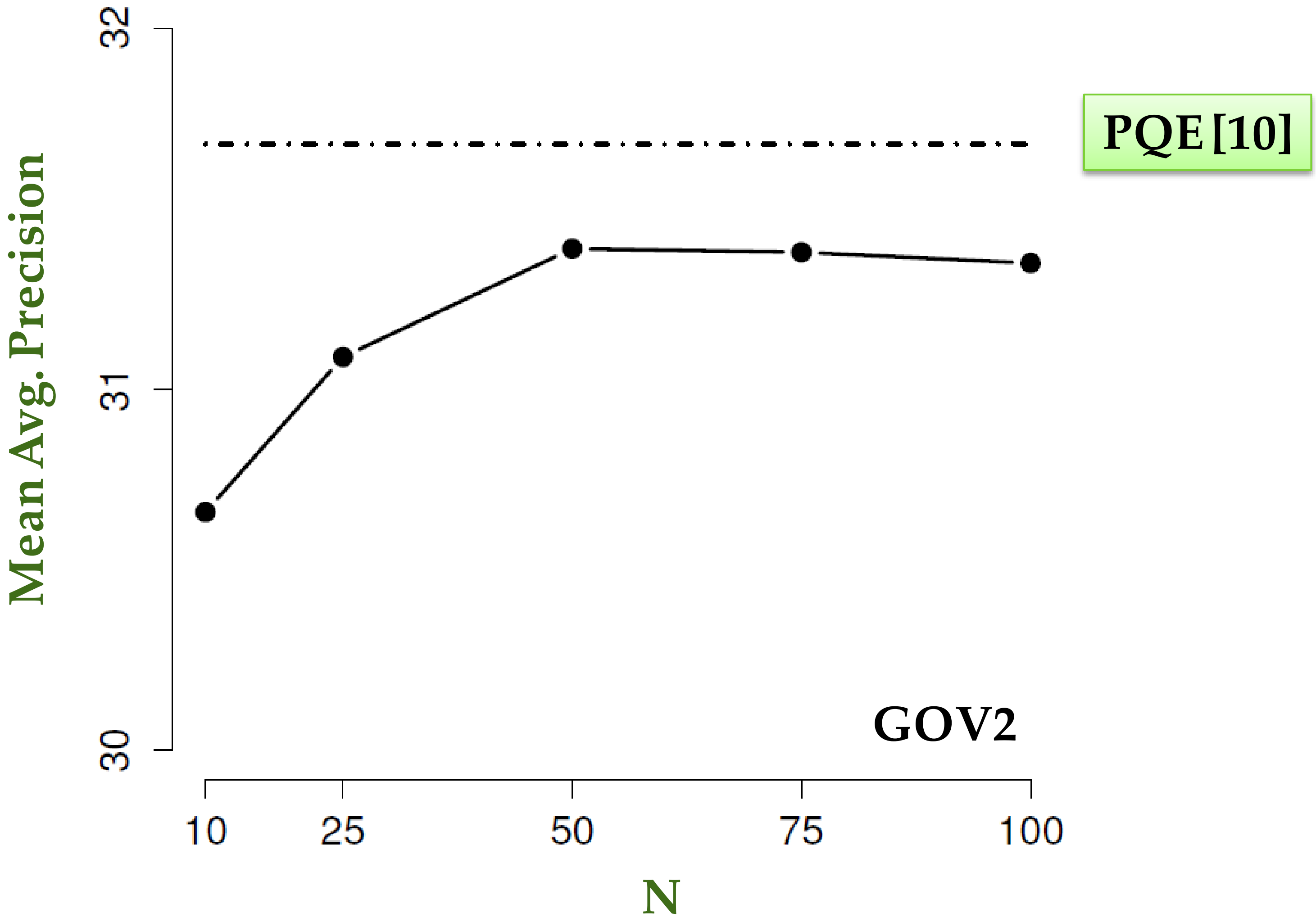
Comparison with Expansion & Weighting Methods

	Query Terms	Exact Phrases	Proximity Matches	Expansion Terms
Weighted Sequential Dependence **(WSD)**	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	
Latent Concept Expansion **(LCE)**	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$	$\mathcal{N}$
Parameterized Query Expansion **(PQE)**	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$	$\mathcal{P}$

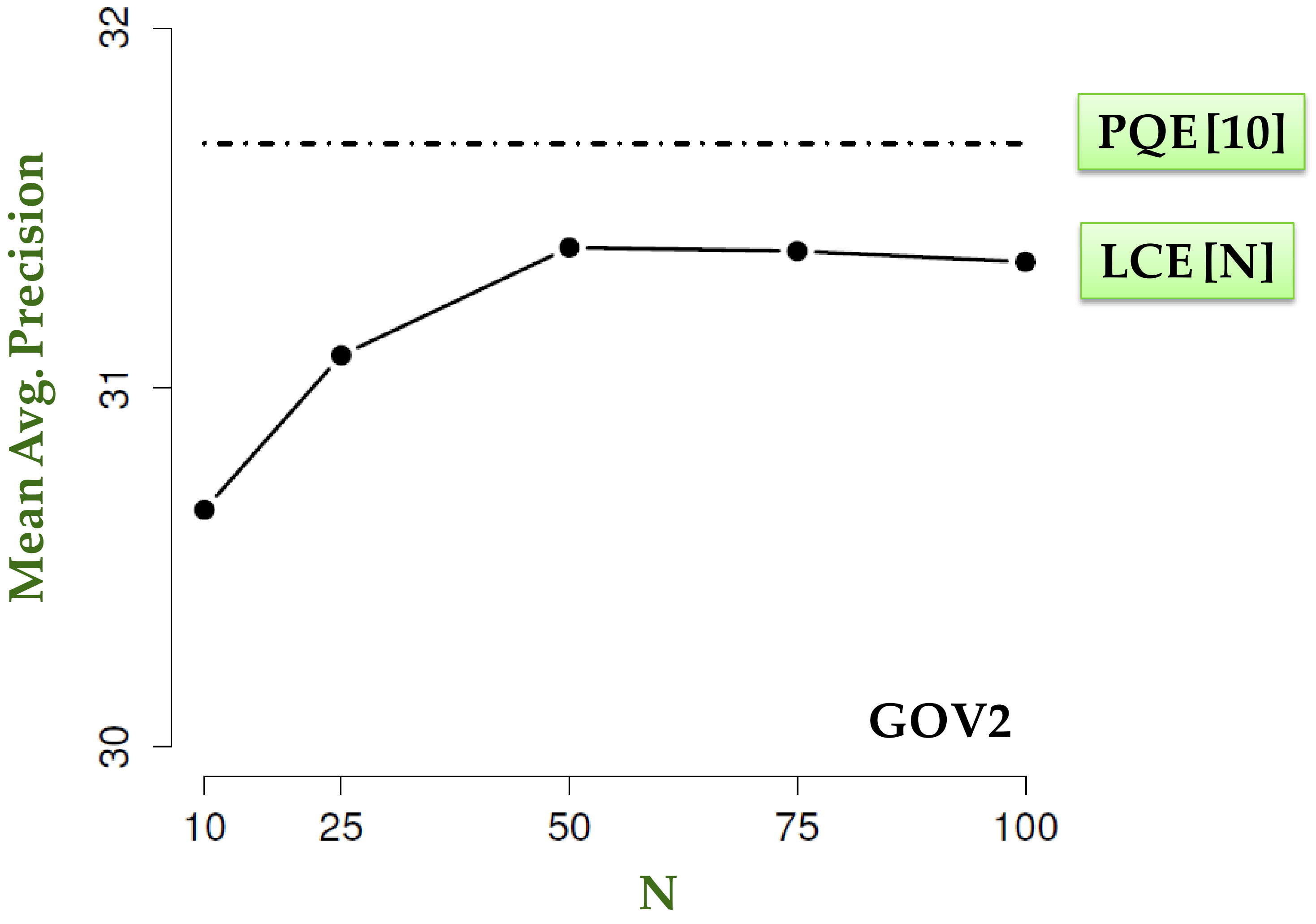
Comparison with Expansion & Weighting Methods



Number of Expansion Terms



Number of Expansion Terms



1. Search Query Representation
2. Parameterized Concept Weighting
3. Explicit Concept Weighting
4. Expansion Concept Weighting
5. **Concept Weighting on Web Scale**

CONCEPT WEIGHTING ON WEB SCALE

(Bendersky et. al, in submission)

Expansion & Weighting Challenges on Web Scale

- **Large variance in web page quality**
  - *Noisy collection statistics*
  - *Noisy expansion terms*
- Need for succinct queries
  - *Minimal query expansion*
- Efficient concept weighting & expansion

Expansion & Weighting Challenges on Web Scale

- Large variance in web page quality
  - *Noisy collection statistics*
  - *Noisy expansion terms*

- **Need for succinct queries**
  - *Minimal query expansion*

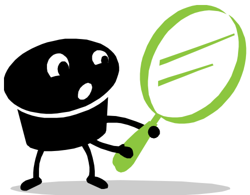
- Efficient concept weighting & expansion

Expansion & Weighting Challenges on Web Scale

- Large variance in web page quality
  - *Noisy collection statistics*
  - *Noisy expansion terms*

- Need for succinct queries
  - *Minimal query expansion*

- **Efficient concept weighting & expansion**



ER TV show

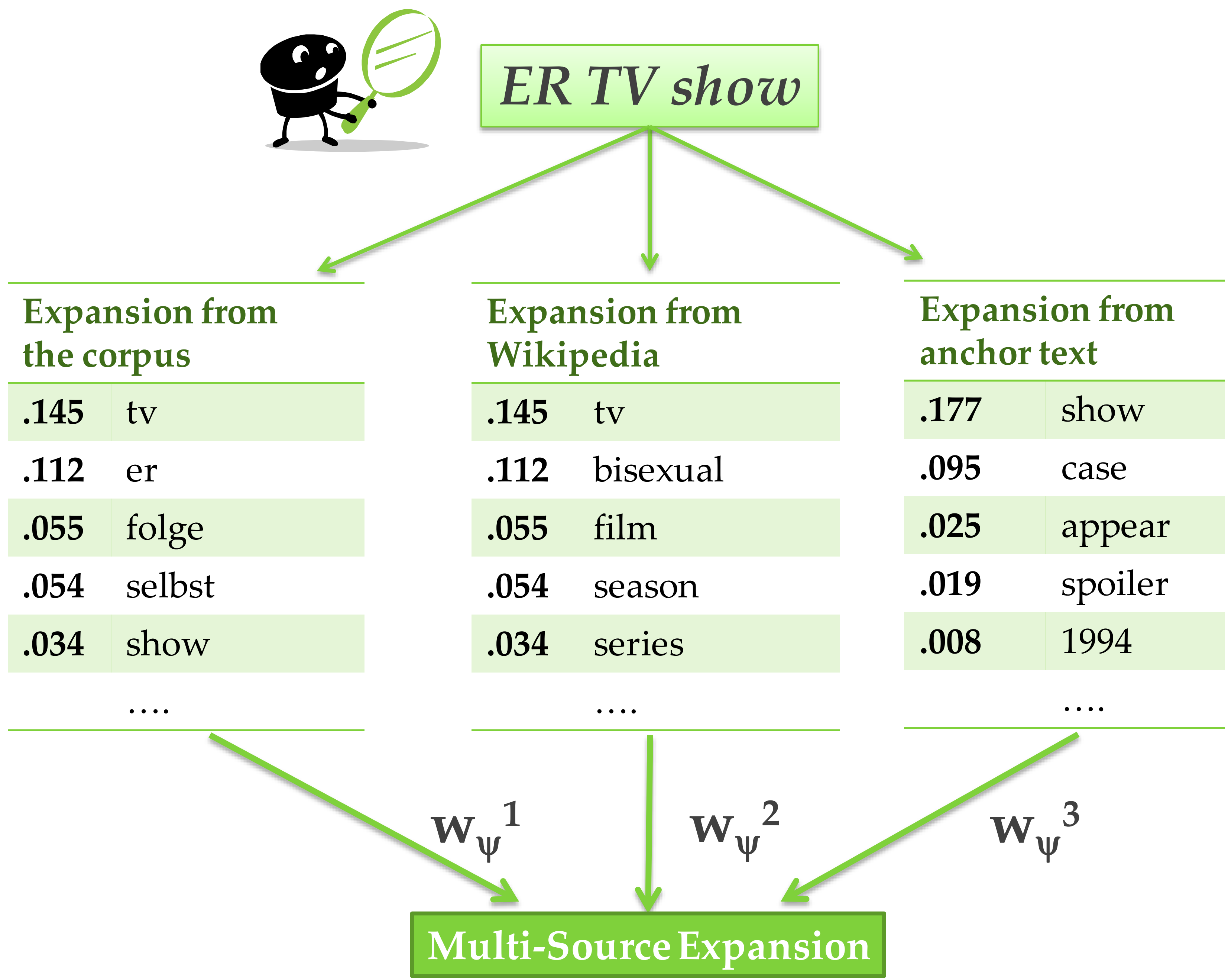
Expansion from the corpus	
.145	tv
.112	er
.055	folge
.054	selbst
.034	show
	....

MAP = 12.29

ER TV show

Expansion from Wikipedia	
.145	tv
.112	bisexual
.055	film
.054	season
.034	series
	….

MAP = 25.68



ER TV show
Expansion from the corpus
.145 tv
.112 er
.055 folge
.054 selbst
.034 show
....
Expansion from Wikipedia
.145 tv
.112 bisexual
.055 film
.054 season
.034 series
....
Expansion from anchor text
.177 show
.095 case
.025 appear
.019 spoiler
.008 1994
....
w_ψ^1
w_ψ^2
w_ψ^3
Multi-Source Expansion

ER TV show

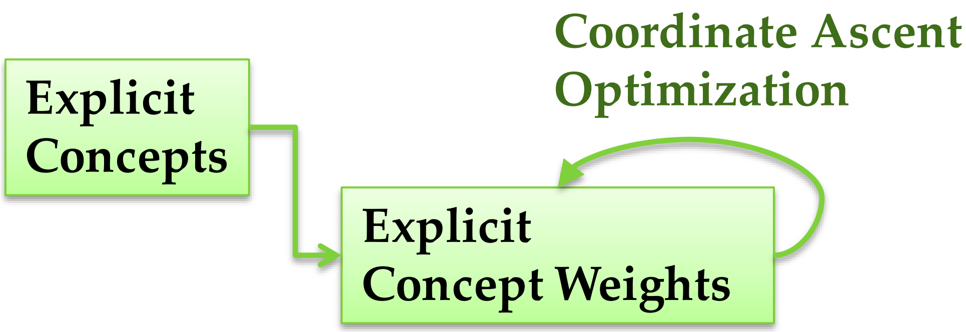
Multi-Source Expansion	
.085	season
.065	episode
.051	dr
.043	drama
.036	series
	….

MAP = 38.31

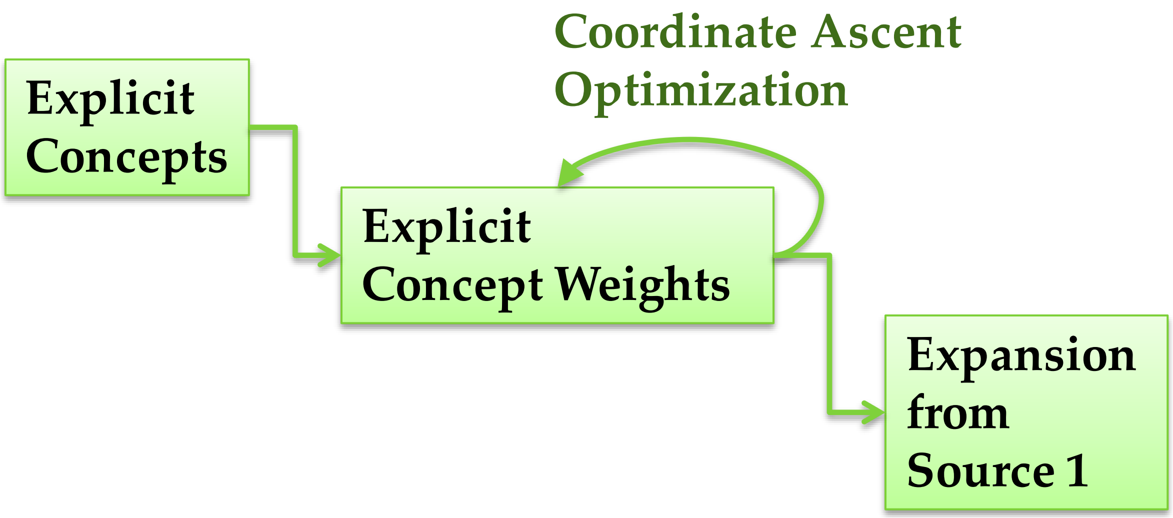
Summary of External Sources

External Source	Description
Web Headings	Text in the <h*> tags in HTML mark-up
Anchor Text	Text in the <a> tag in HTML mark-up
Wikipedia Corpus	Wikipedia articles
Retrieval Corpus	Large web collection (ClueWeb)

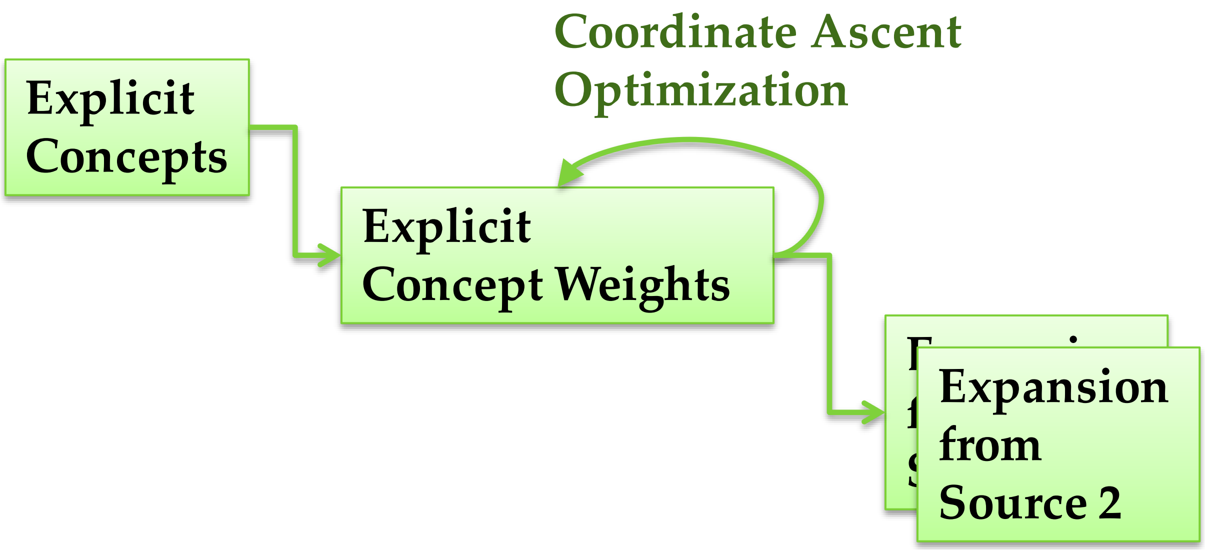
MSE Training Process



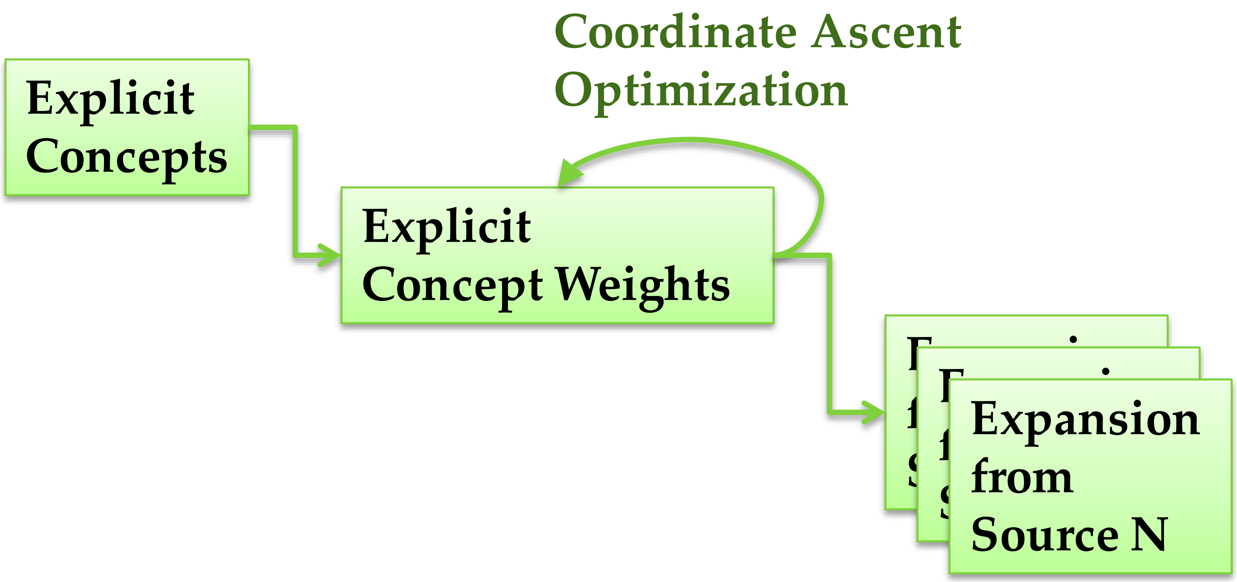
MSE Training Process



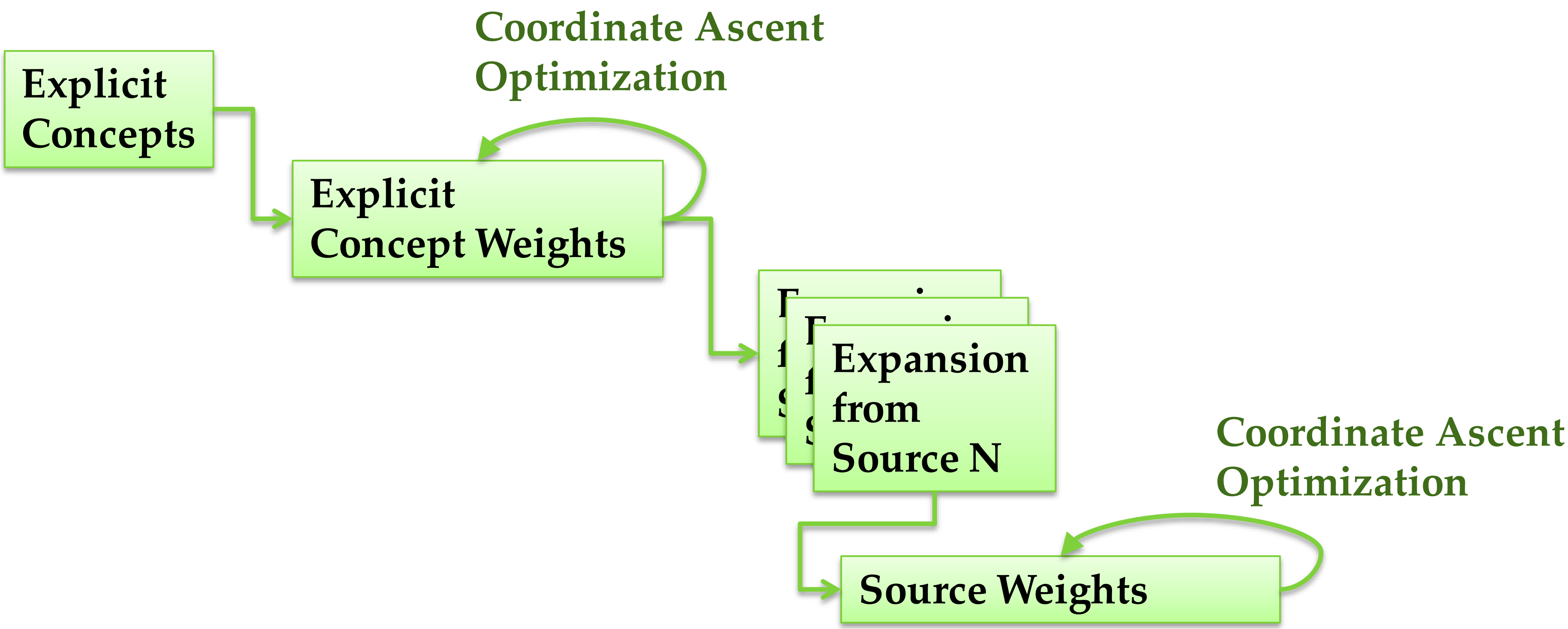
MSE Training Process



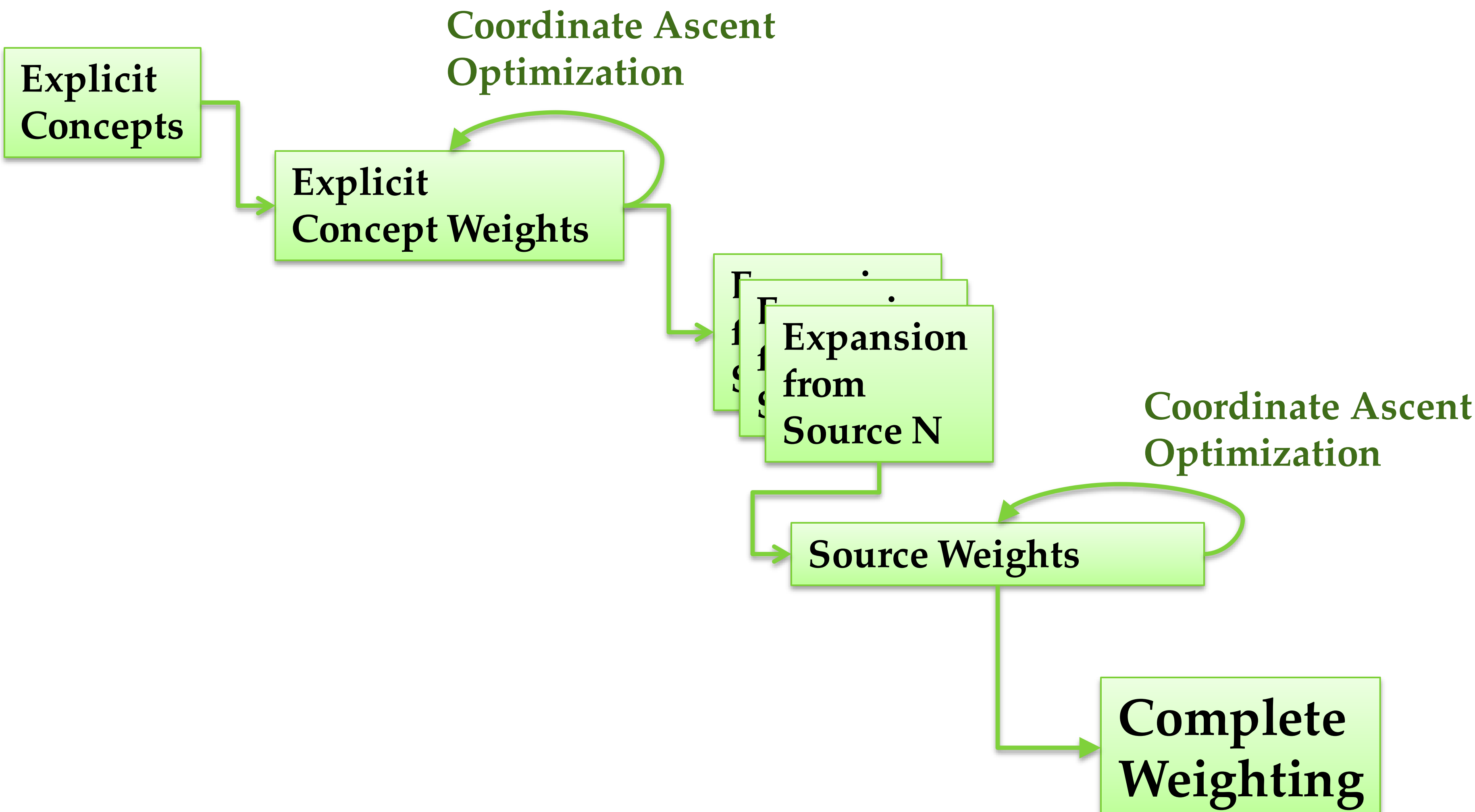
MSE Training Process



MSE Training Process



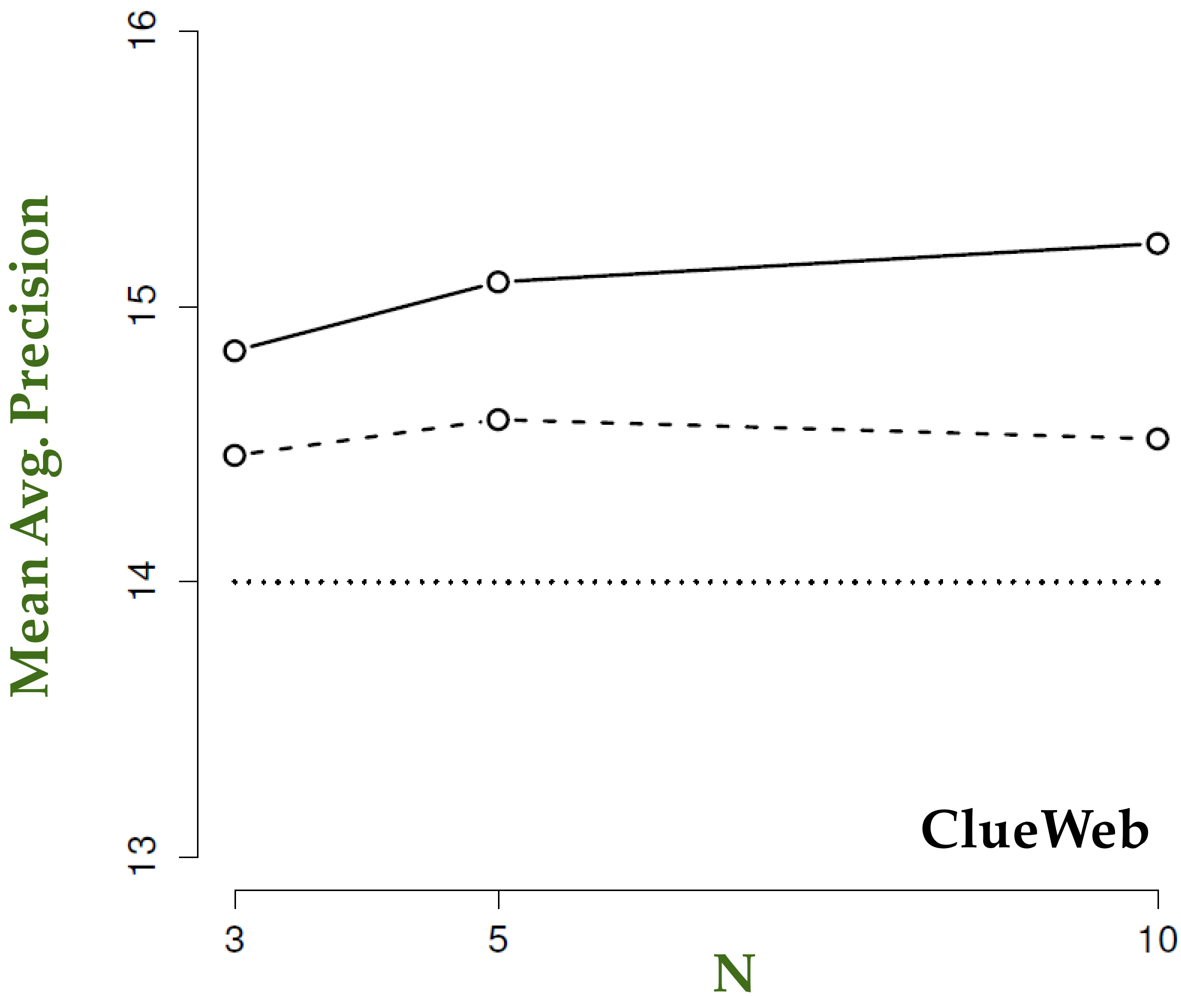
MSE Training Process



Comparison with Parameterized Methods

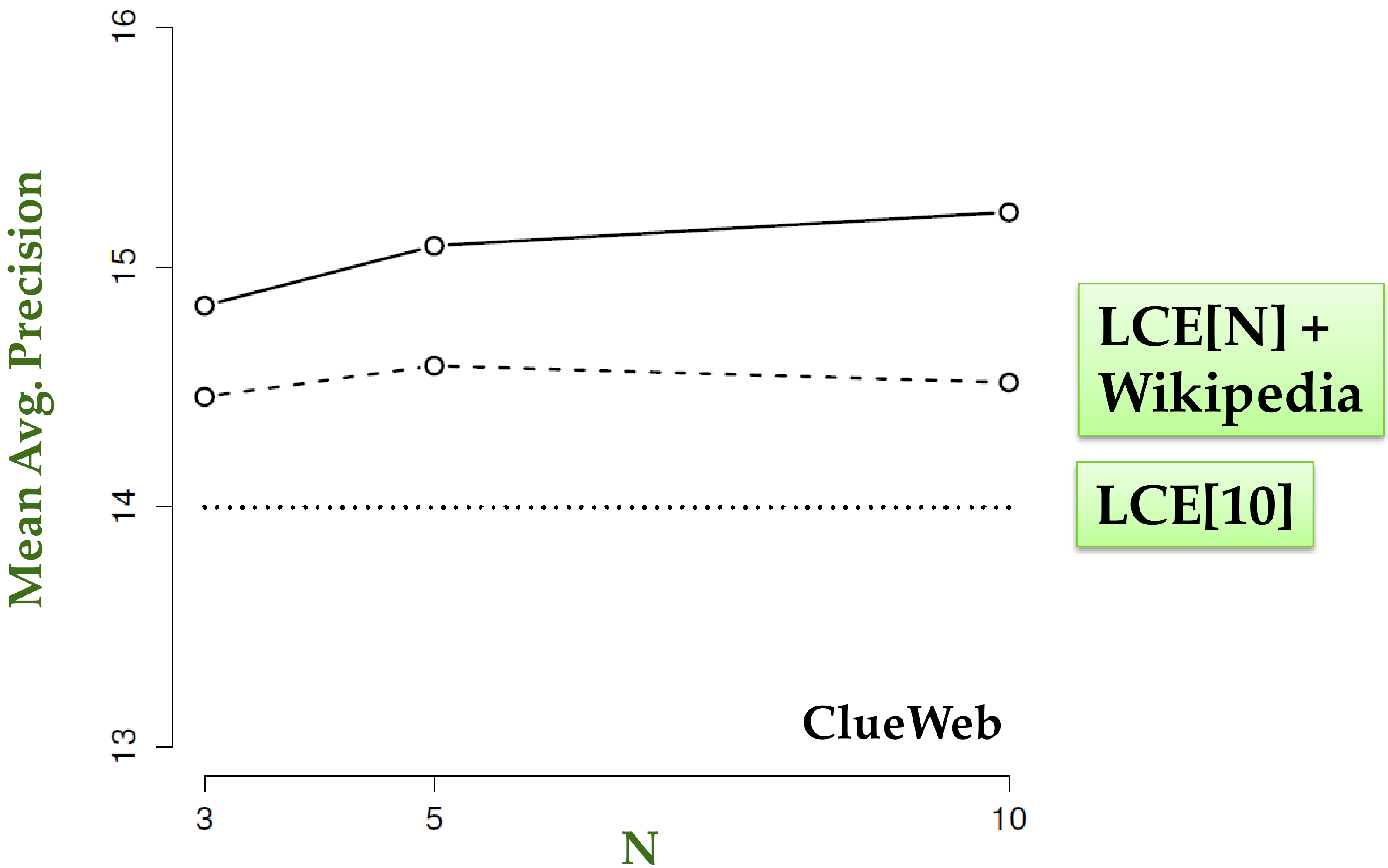
Large Web Corpus

Number of Expansion Terms

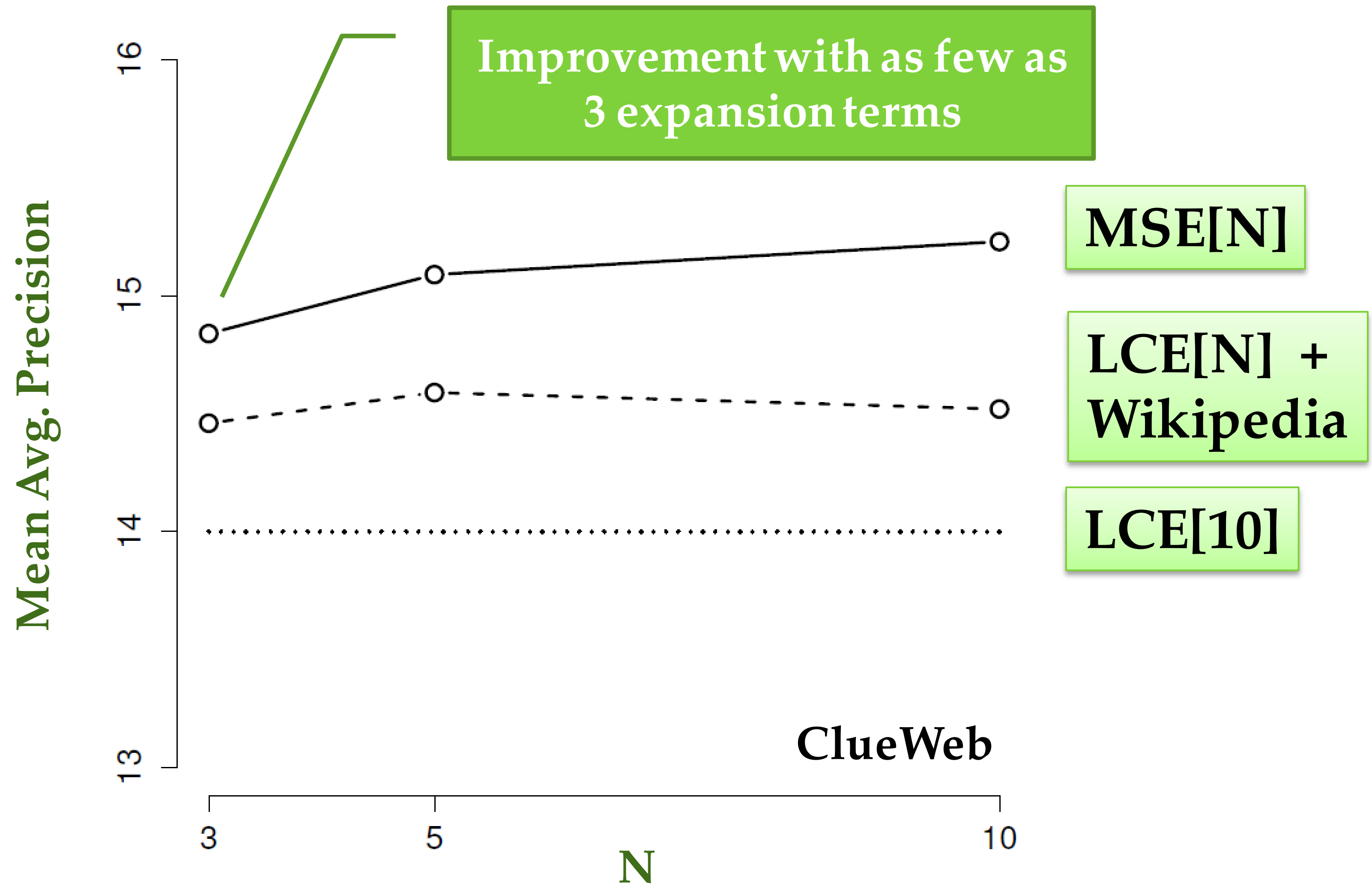


LCE[10]

Number of Expansion Terms

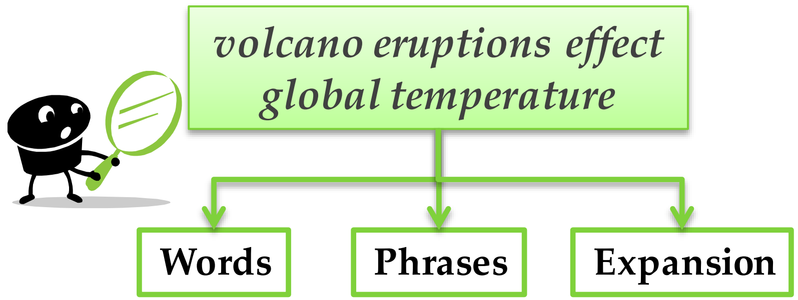


Number of Expansion Terms



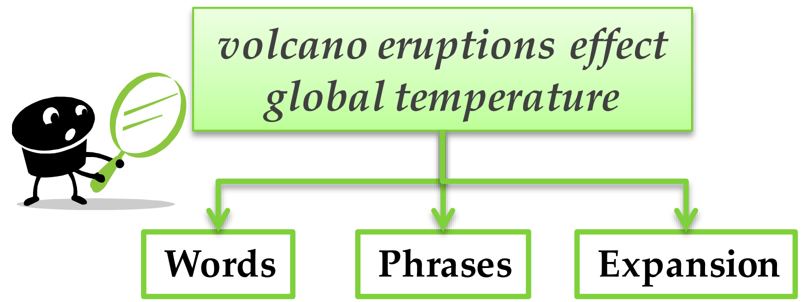
SUMMARY

Query Representation – Important Research Problem



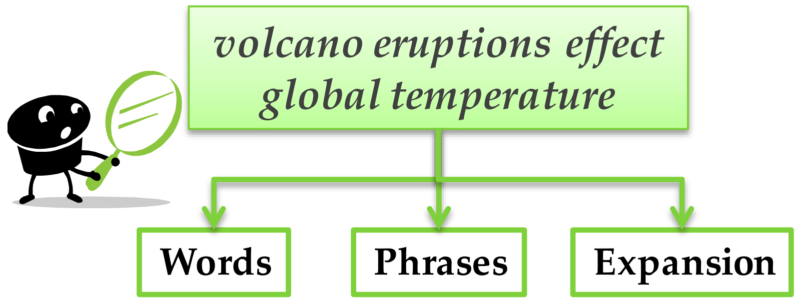
Impacts billions of search queries

Query Representation – Important Research Problem



Improves understanding of user search behavior

Query Representation – Important Research Problem



Synthesis of ideas

Information Retrieval
Natural Language Processing
Machine Learning

Query Representation & Understanding Workshop

http://ciir.cs.umass.edu/sigir2011/qru/

- Short research papers & invited talks
- SIGIR Forum publication
- New public dataset

Parameterized Concept Weighting

- **Novel information retrieval framework**
- More realistic modeling of user intent compared to previous work
- Significant gains in effectiveness compared to current state-of-the-art IR models

Parameterized Concept Weighting

- Novel information retrieval framework

- **More realistic modeling of user intent compared to previous work**

- Significant gains in effectiveness compared to current state-of-the-art IR models

Parameterized Concept Weighting

- Novel information retrieval framework
- More realistic modeling of user intent compared to previous work
- **Significant gains in effectiveness compared to current state-of-the-art IR models**

More to Come...

- **More complex query representations**

- **Integration with web-scale ranking systems**
  - *Scaling to hundreds/thousands features*

- **Applications in other domains**
  - *Q&A systems*
  - *Content Matching & Recommendation*

More to Come...

- More complex query representations
- **Integration with web-scale ranking systems**
  - *Scaling to hundreds/thousands features*
- Applications in other domains
  - *Q&A systems*
  - *Content Matching & Recommendation*

More to Come...

- **More complex query representations**
- **Integration with web-scale ranking systems**
  - *Scaling to hundreds/thousands features*
- **Applications in other domains**
  - *Q&A systems*
  - *Content Matching & Recommendation*

THANK YOU!