Improving IR performance from OCRed text using cooccurrence

RISOT 2012

Kripabandhu Ghosh
and
Anirban Chakraborty
Indian Statistical Institute
Kolkata, India

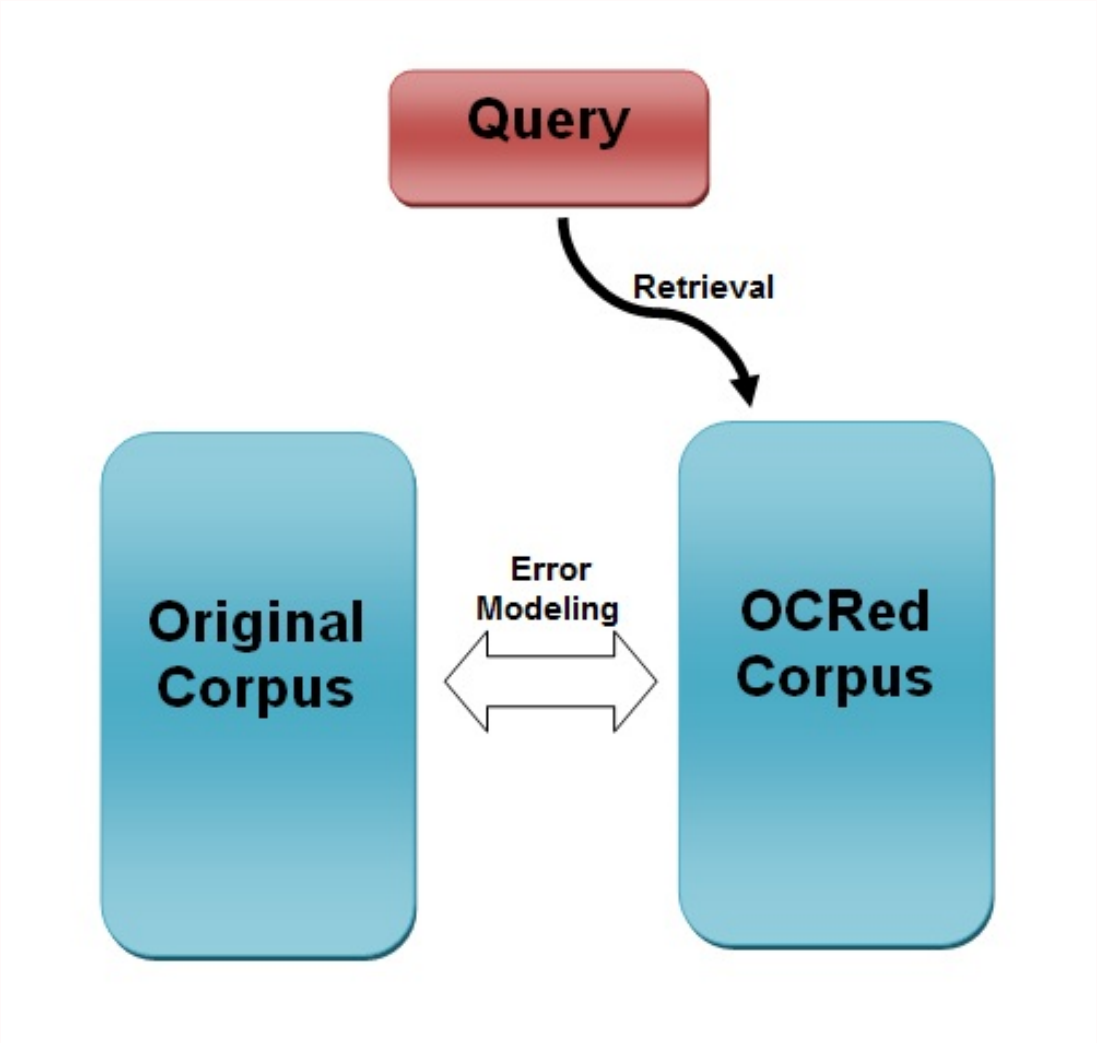
Title

Improving IR performance from OCRed text using cooccurrence RISOT 2012

Key Terms

- Co-occurrence : We say that two words co-occur if they appear in a window of certain number of words of each other in a document
- LCS_similarity : **LCS** stands for **L**ongest **C**ommon **S**ubsequence
  - LCS(*industry*, *industrial*) = *industr*
  - LCS_similarity(*industry*, *industrial*) = $|\text{LCS}(\textit{industry}, \textit{industrial})|/\max(|\textit{industry}|, |\textit{industrial}|) = 0.7$

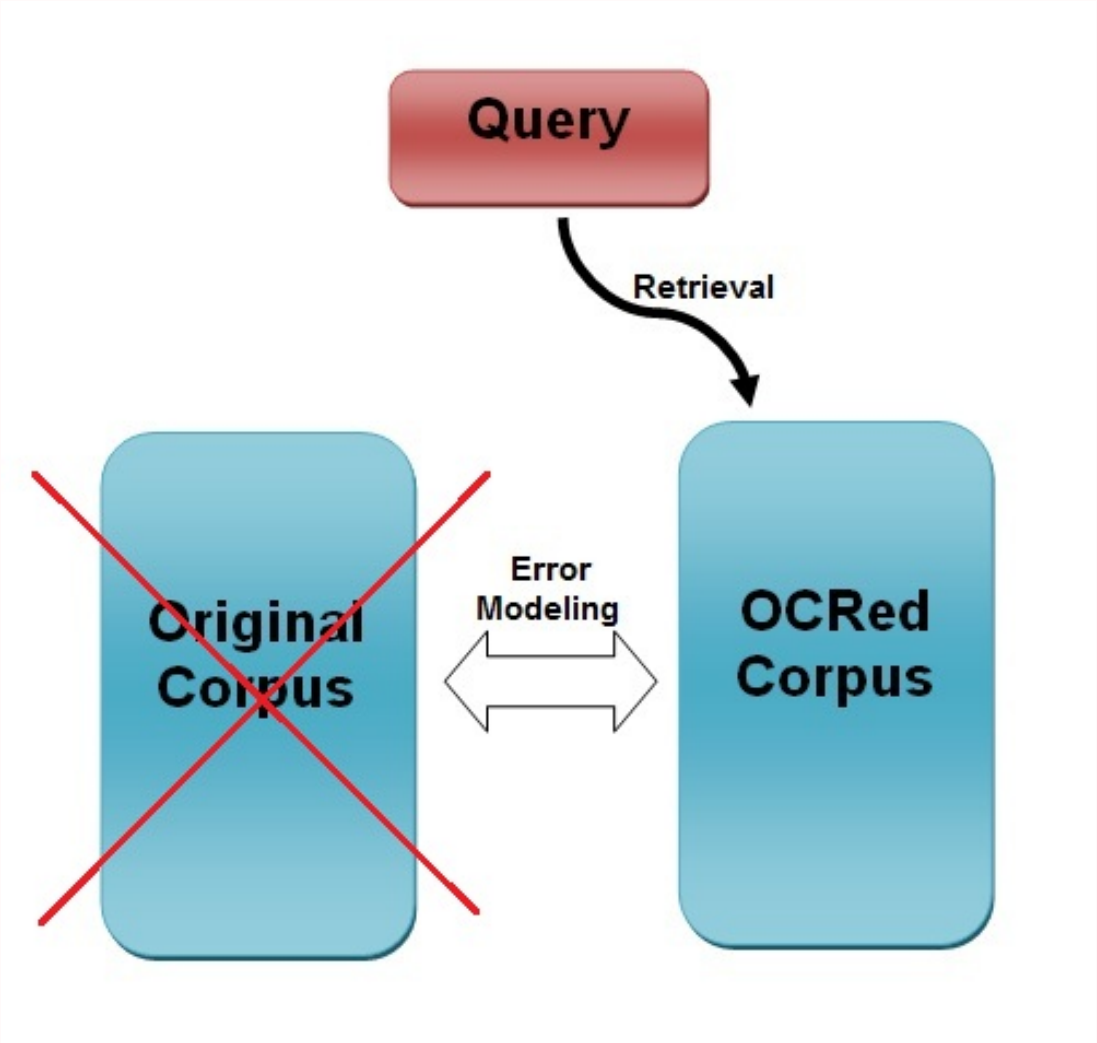
RISOT task


Query
Retrieval
Original Corpus
Error Modeling
OCRed Corpus

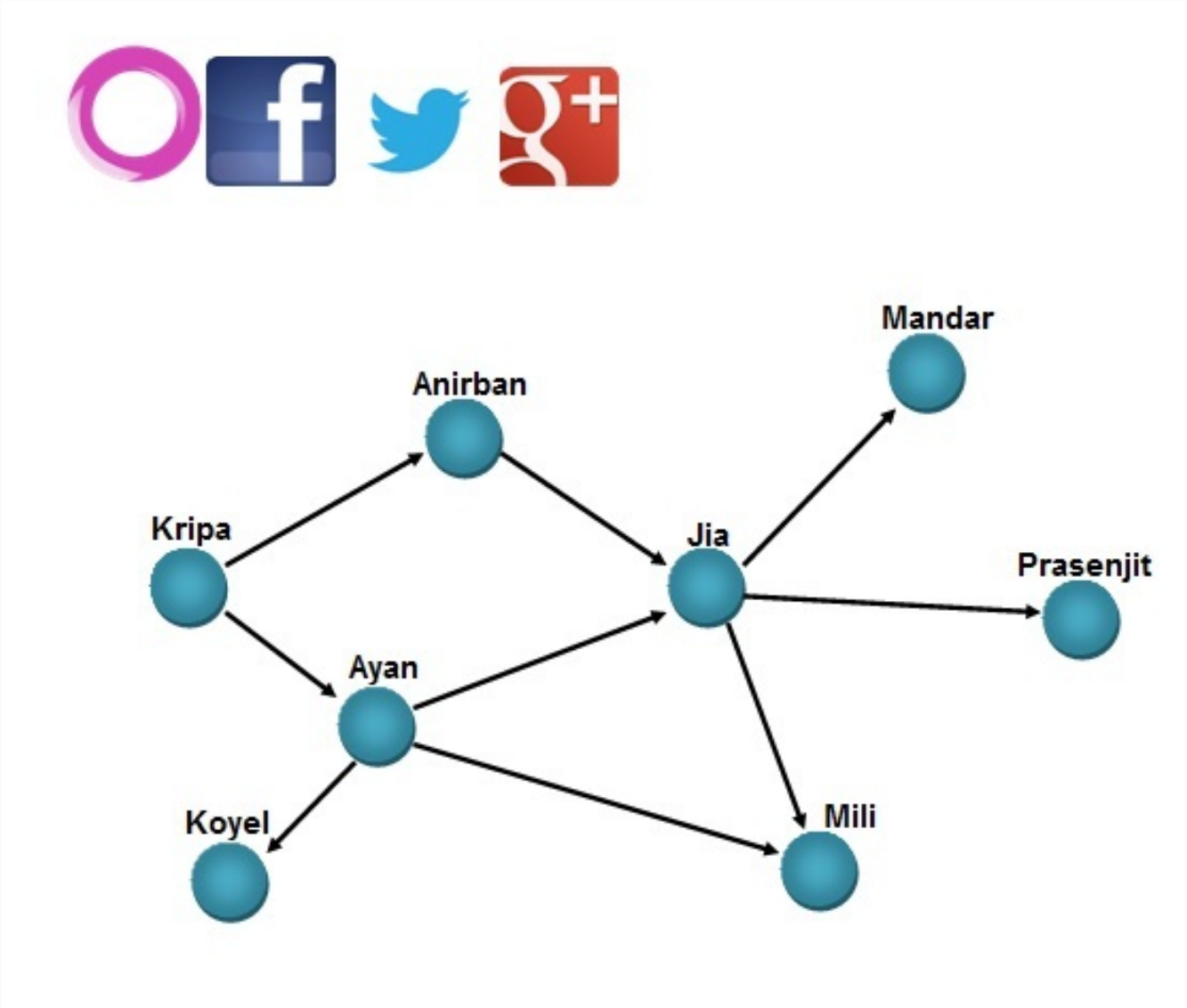
OCRed Resources

- Legal documents as hard copies
  - IIT CDIP 1.0 corpus - TREC Legal Ad Hoc
- SIGIR Digital Museum - Cleverdon, Salton, Sparck Jones

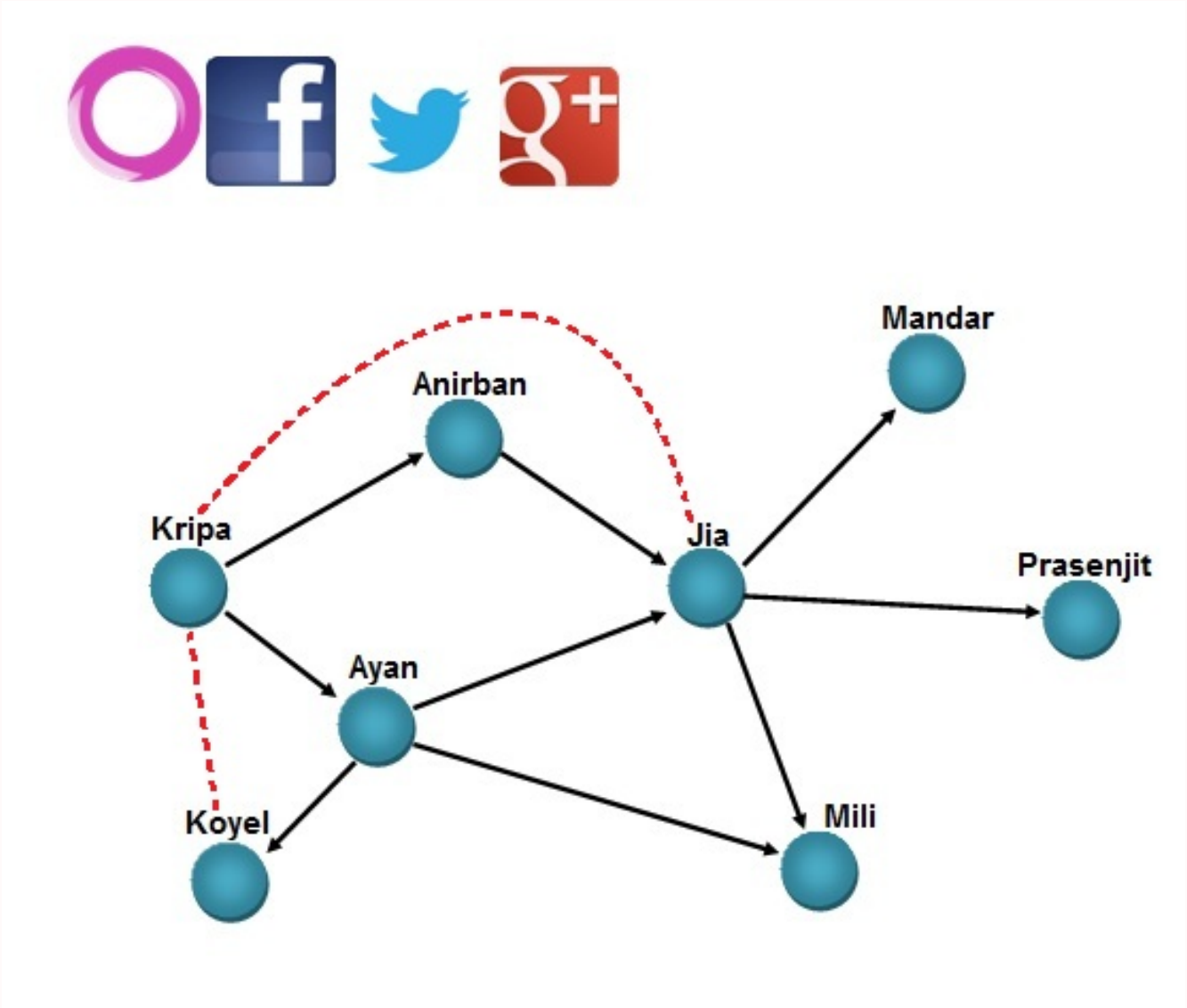
RISOT task - Without Original Corpus



Social Networks : Direct Connections



Social Networks : Indirect Connections



Clustering Algorithm

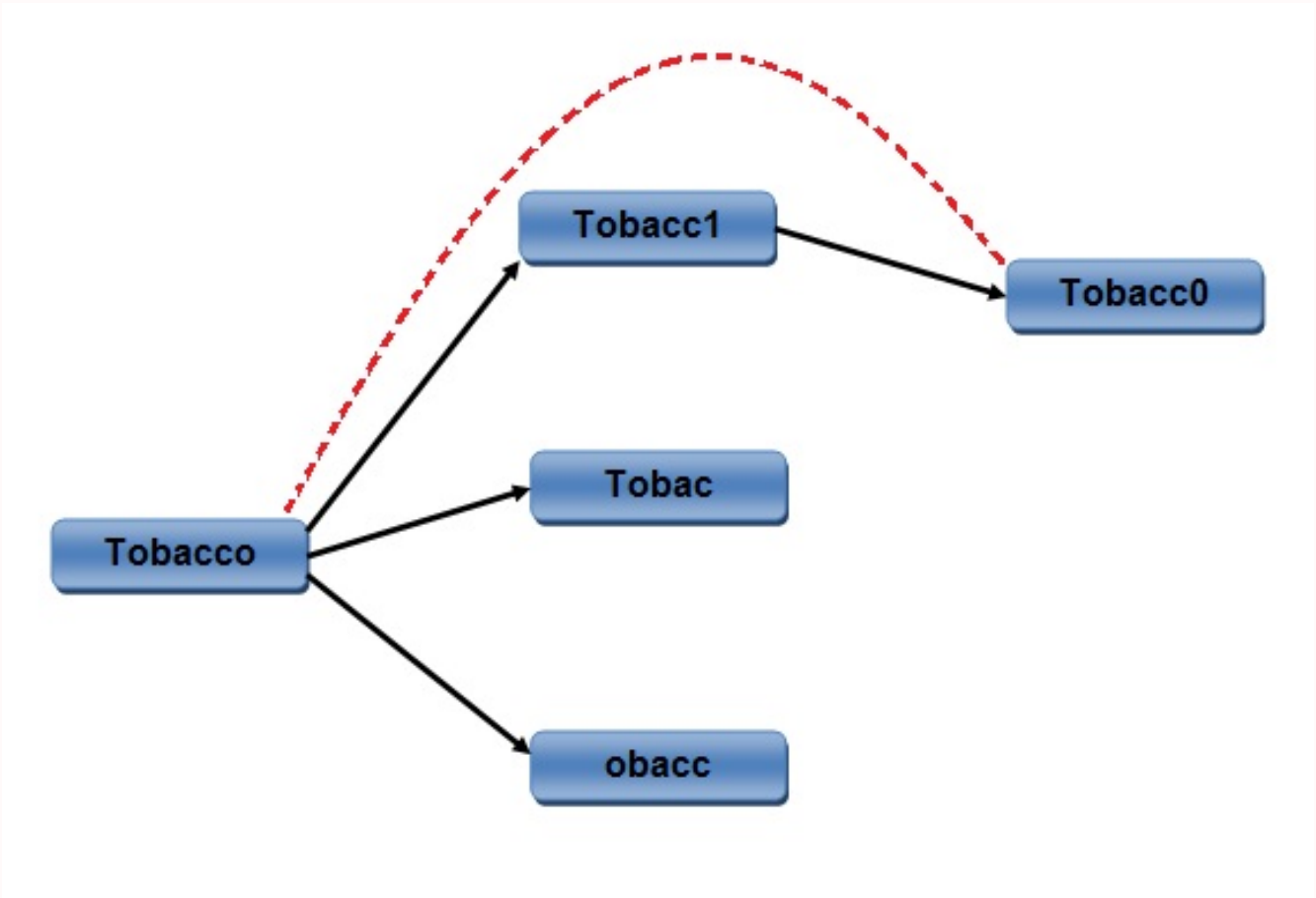
CLUSTERING

Clustering Algorithm (Phase I)

For each word w in the OCRed corpus:
Let S_{w1} and S_{w2} be empty sets. Let $S_w = S_{w1} \cup S_{w2}$

1. For word w_1 co-occurring with w, calculate LCS_similarity between w and w_1. Store w_1 in S_{w1} if LCS_similarity(w, w_1) $>$ some threshold T.
2. For each w' in S_{w1}, find the words w_2 co-occurring with w' such that LCS_similarity(w, w_2) $>$ T. Include all these words in S_{w1}.
3. Repeat step (2) until no new word is added to S_{w1}.

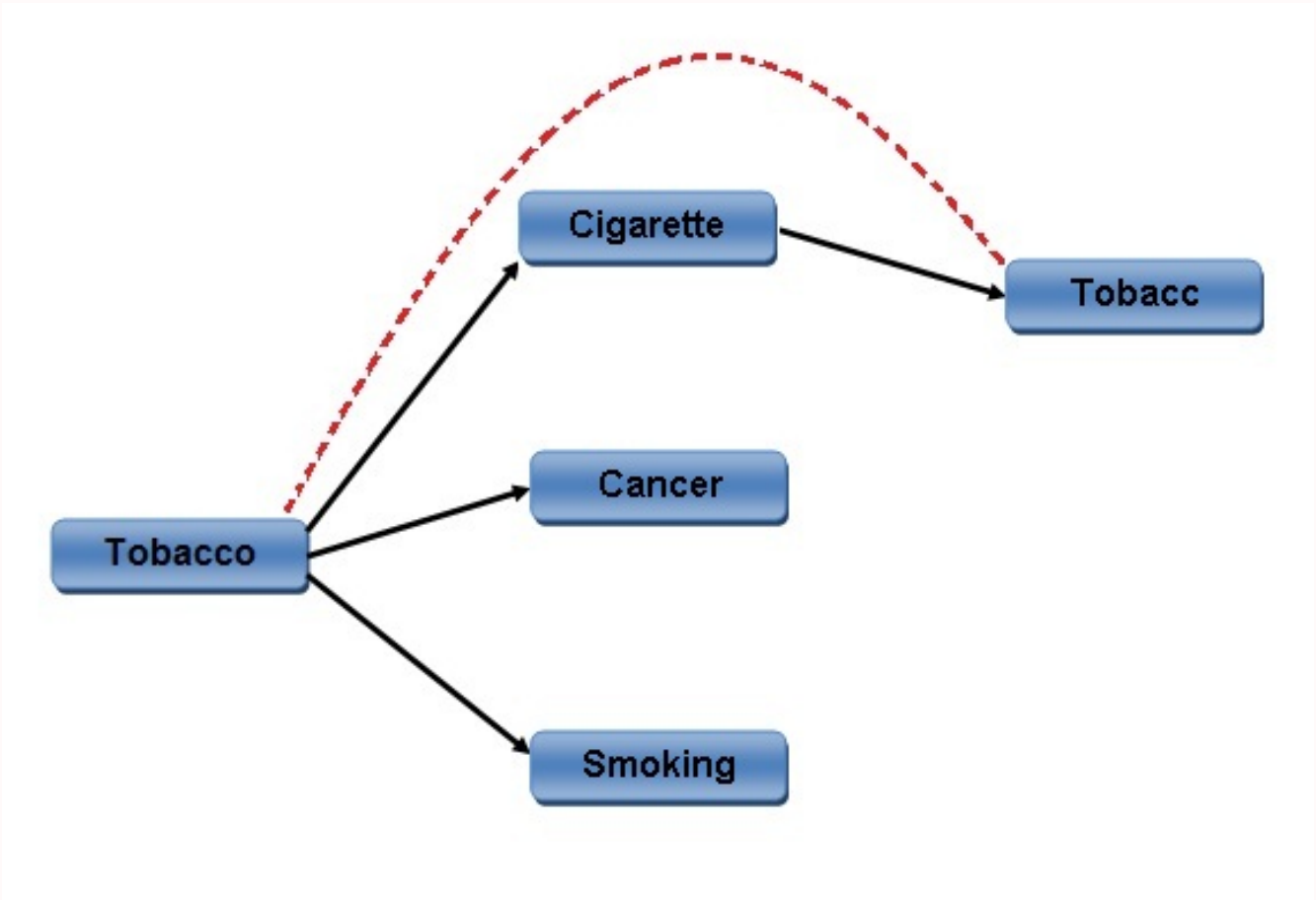
Clustering Algorithm (Phase I)



Clustering Algorithm (Phase II)

1. Consider top m (in terms of frequency in corpus) words co-occurring with w. For each such word w_3, find the words w_4 cooccurring with w_3 such that LCS_similarity(w, w_4) $>$ T. Include all these words in S_{w2}.
2. For each w'' in S_{w2}, find the words w_5 co-occurring with w'' such that LCS_similarity(w, w_5) $>$ T. Include all these words in S_{w2}.
3. Repeat step (2) until no new word is added to S_{w2}.

Clustering Algorithm (Phase II)



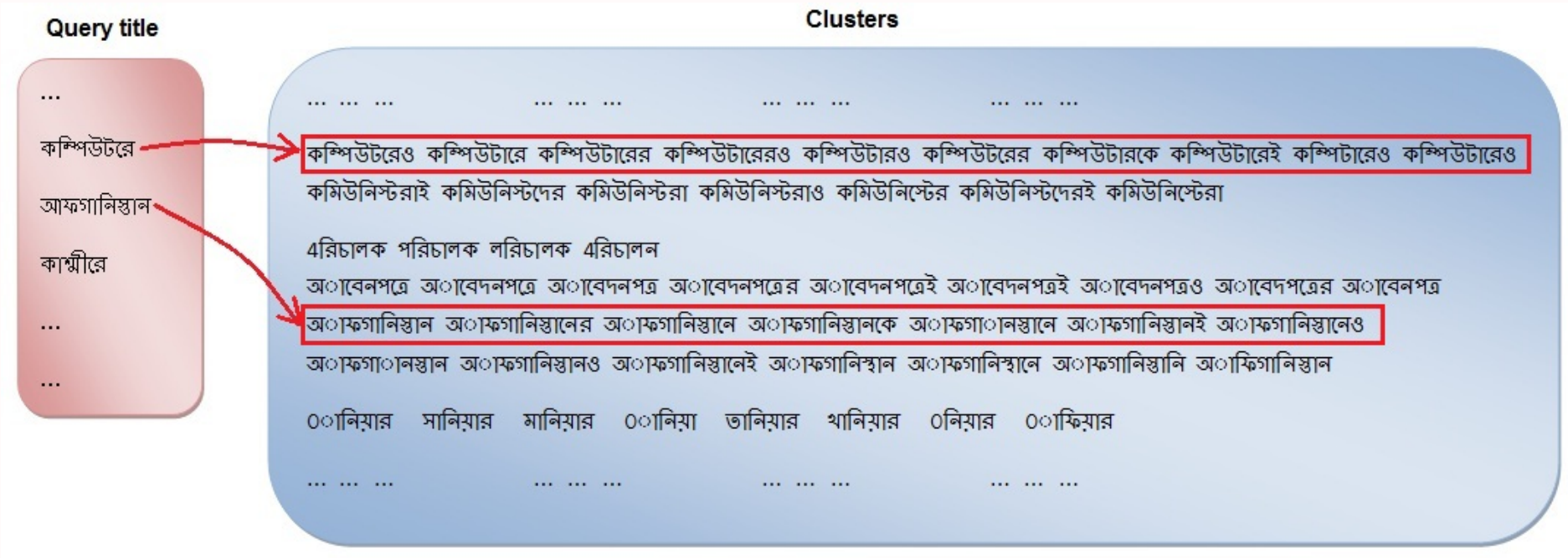
Query word cluster mapping

1. calculate LCS_similarity(w_q, w_C), where w_C is a word in cluster C, for each word in the corpus
2. Choose all the clusters C for which LCS_similarity(w_q, w_C) is greater than a high threshold
3. For each cluster C obtained from step (2) define $C' = C \cup \{w_q\}$
4. Create complete-linkage clusters from each C' of step (3) and keep those clusters containing w_q

Query word cluster mapping

5. For a cluster C, let us consider LCS_similarity between each pair of words in it. Let GM_C denote the Geometric Mean of LCS_similarity of all the pairs. Then, compute GM_C for each cluster given by step (4)
6. Select the cluster C with maximum GM_C as the appropriate cluster for w_q

Query word clusters

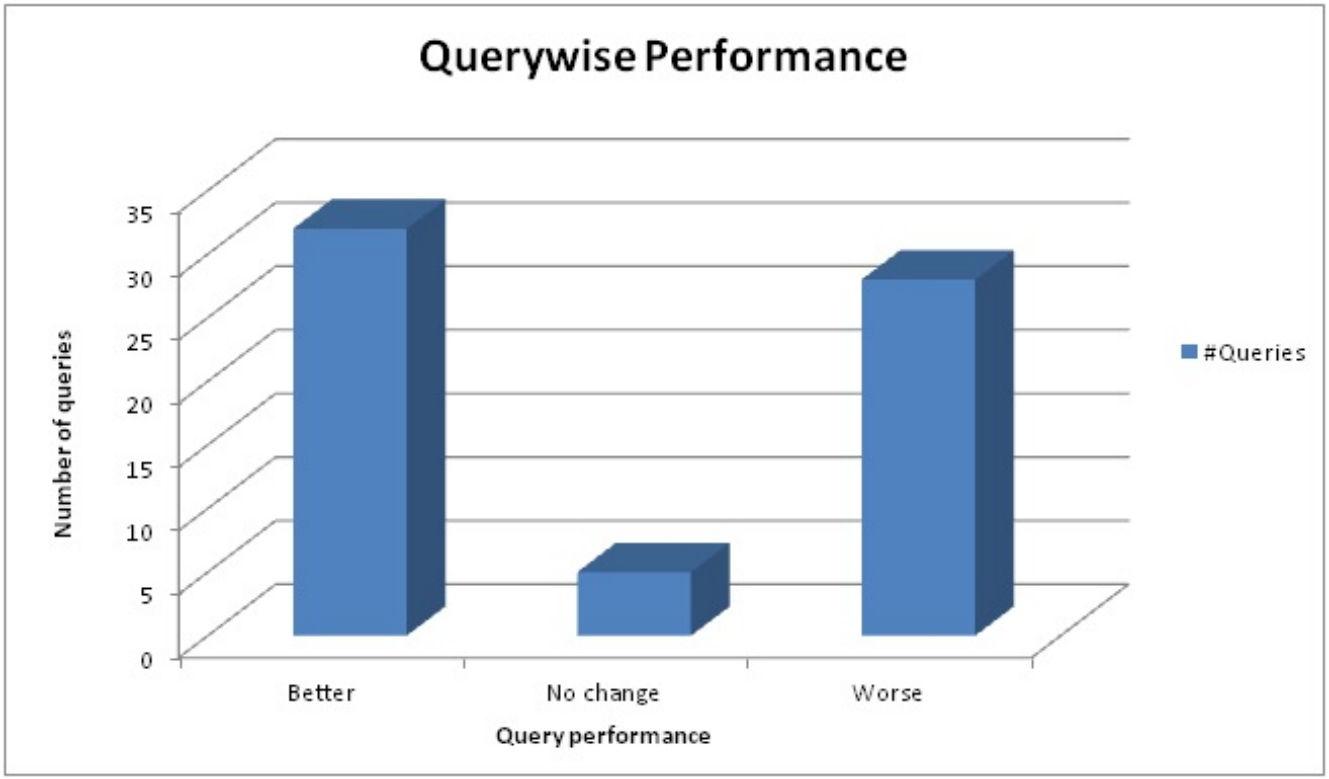


Results

Run	MAP	P5
Original text	0.2567	0.3485
OCRed text (baseline)	0.1791	0.2738
Proposed method on OCRed text	0.1974[1]	0.2831

[1] Not significant

Querywise Performance



Failure Analysis : clusters

The Good

... ...

কম্পিউটরেও কম্পিউটারে কম্পিউটারের কম্পিউটারেরও কম্পিউটারও কম্পিউটরের কম্পিউটারকে কম্পিউটারেই কম্পিটারেও কম্পিউটারেও
কমিউনিস্টরাই কমিউনিস্টদের কমিউনিস্টরা কমিউনিস্টরাও কমিউনিস্টের কমিউনিস্টদেরই কমিউনিস্টেরা
অাবেনপত্রে অাবেদনপত্রে অাবেদনপত্র অাবেদনপত্রের অাবেদনপত্রেই অাবেদনপত্রই অাবেদনপত্রও অাবেদপত্রের অাবেনপত্র
অাফগানিস্তান অাফগানিস্তানের অাফগানিস্তানে অাফগানিস্তানকে অাফগাানস্তানে অাফগানিস্তানই অাফগানিস্তানেও অাফগাানস্তান
অাফগানিস্তানও অাফগানিস্তানেই অাফগানিস্থান অাফগানিস্থানে অাফগানিস্তানি অাফিগানিস্তান

... ...

and...

The Bad

... ...

4রিচালক পরিচালক লরিচালক 4রিচালন
০ানিয়ার সানিয়ার মানিয়ার ০ানিয়া তানিয়ার খানিয়ার ০নিয়ার ০াফিয়ার

... ...

Failure Analysis

- Compact clusters over all-inclusive clusters
- Re-clustering based on string match and co-occurrence
- Chance co-occurrence - harmful
- Incorporation of co-occurrence frequencies - essential

Brighter side

- Practical utility
- Language independent
- Context information - reliable
- Captures both erroneous and inflectional variants (effect of stemming)

THANK YOU