A System for Query-Specific Document Summarization

Ramakrishna Varadarajan,

Vagelis Hristidis.

FLORIDA INTERNATIONAL UNIVERSITY,
School of Computing and Information Sciences,
Miami.

Roadmap

- Need for query-specific summaries
- Our approach
  - Building a document graph
  - Definition of summary
  - Rank Summaries
- Efficient computation of summaries
- Evaluation of summarization process
  - Quality
  - Performance
- Related Work
- Conclusions

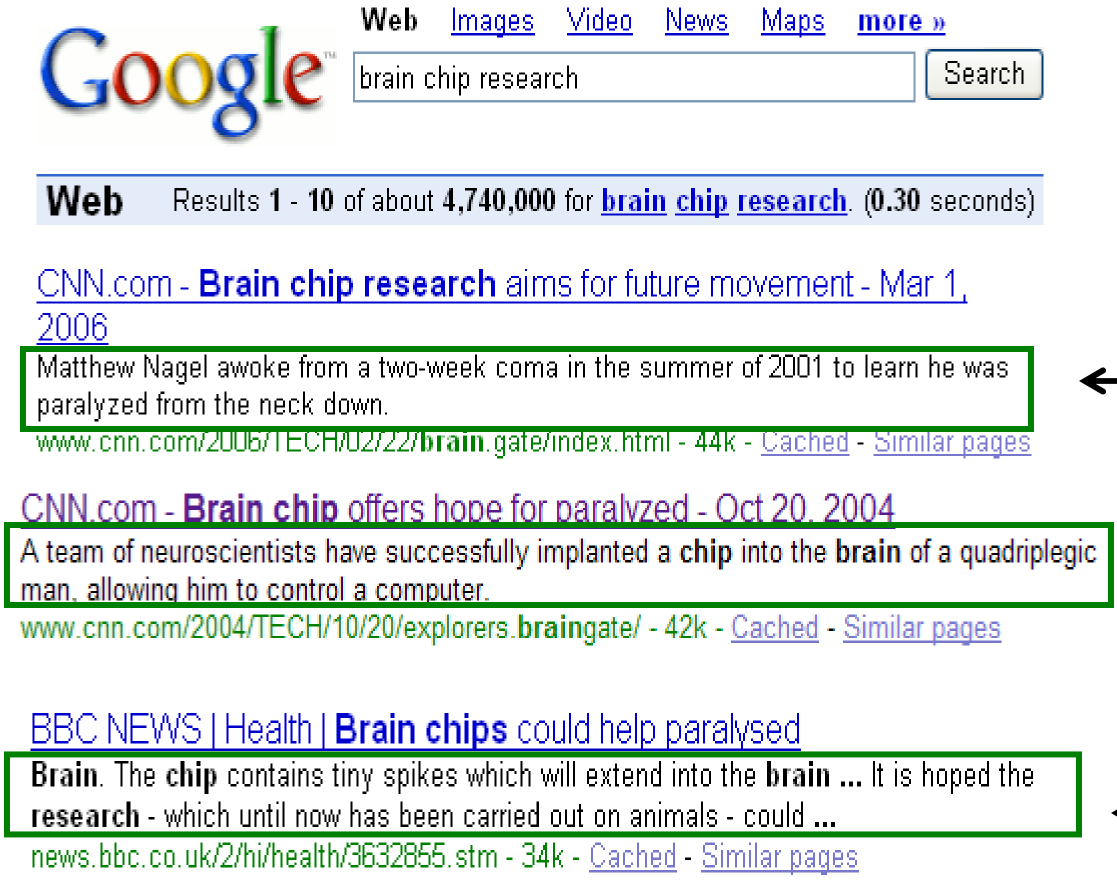
Roadmap

- Need for query-specific summaries
- Our approach
  - Building a document graph
  - Definition of summary
  - Rank Summaries
- Efficient computation of summaries
- Evaluation of summarization process
  - Quality
  - Performance
- Related Work
- Conclusions

Need for Query-Specific Summaries

- Locating relevant information is hard.
- **Summaries** are **helpful** because:
  - Provide a Quick preview of the document.
  - Allow users to quickly decide relevance.
  - Save user’s browsing time.
- Success of *Web search engines* – Query specific **snippets** are important.
- Two categories of summaries:
  - *Query-Independent* – Most of prior works.
  - *Query-Specific* – Applicable to web search engines.

Motivation



brain chip research

Web Results **1** - **10** of about **4,740,000** for **brain chip research**. (**0.30** seconds)

CNN.com - **Brain chip research** aims for future movement - Mar 1, 2006

Matthew Nagel awoke from a two-week coma in the summer of 2001 to learn he was paralyzed from the neck down.

www.cnn.com/2006/TECH/02/22/**brain**.gate/index.html - 44k - Cached - Similar pages

CNN.com - **Brain chip** offers hope for paralyzed - Oct 20, 2004

A team of neuroscientists have successfully implanted a **chip** into the **brain** of a quadriplegic man, allowing him to control a computer.

www.cnn.com/2004/TECH/10/20/explorers.**braingate**/ - 42k - Cached - Similar pages

BBC NEWS | Health | **Brain chips** could help paralysed

Brain. The **chip** contains tiny spikes which will extend into the **brain** ... It is hoped the **research** - which until now has been carried out on animals - could ...

news.bbc.co.uk/2/hi/health/3632855.stm - 34k - Cached - Similar pages

Query-Specific Summaries

Motivation

Drawbacks

- Association between query keywords is unclear.
- Naïve approach for summarization.
- Ignores semantic relations between keywords in the document.

Summarization research till date

- Mostly Query-Independent.
- Not applicable for web search.

Roadmap

- Need for query-specific summaries
- Our approach
  - Building a document graph
  - Definition of summary
  - Rank Summaries
- Efficient computation of summaries
- Evaluation of summarization process
  - Quality
  - Performance
- Related Work
- Conclusions

Our Approach

- Document → ***graph***
- We call it ***Document Graph***.

Three Steps

Step 1: ***Preprocess***

- Build a document graph, *G*.

Step 2: ***Summary Generation***

- Given a query Q and a document graph G,

  Summaries → *Spanning Trees* that cover all keywords

Step 3: ***Rank*** spanning trees.

Building Document Graphs

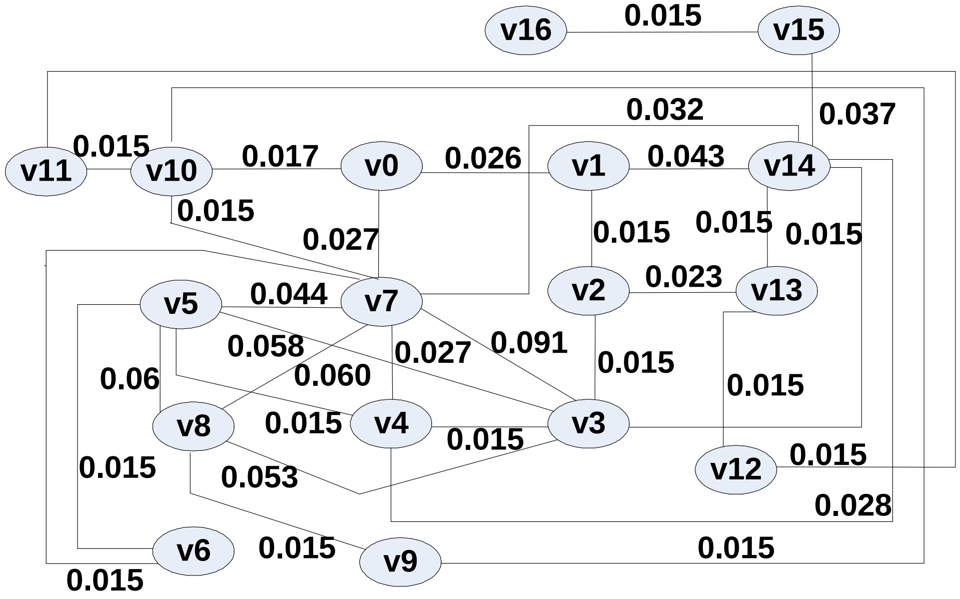
- Parse the document.
- Split it into text fragments (using delimiters or tags).
- Text Fragments represented as *Nodes*
- Add an edge between 2 nodes, if semantically related.
- Edges : Semantic Links
- Edge weights: Degree of association

Example

Sample Document

(v0) Brain chip offers hope for paralyzed
(v1) A team of neuroscientists have successfully implanted a **chip** into the **brain** of a quadriplegic man, allowing him to control a computer.
(v2) …
(v3) …
(v4) …
(v5) BrainGate offers the possibility of hitherto unimaginable levels of independence for the severely disabled.
(v6) …
(v7) …
(v8) …
(v9) …
(v10) Donoghue's initial **research**, published in the science journal Nature in 2002, consisted of attaching an implant to a monkey's **brain** that enabled it to play a simple pinball computer game remotely.
(v11) The four-millimeter square **chip**, which is placed on the surface of the motor cortex area of the **brain**, contains 100 electrodes each thinner than a hair which detect neural electrical activity. The sensor is then connected to a computer via a small wire attached to a pedestal mounted on the skull.
(v12) …
(v13)…
(v14)…
(v15) "Here we have a **research** participant who is capable of controlling his environment by thought alone -- something we have only found in science fiction so far," said Friehs.
(v16) …

Document Graph



- Parsing delimiter – NewLine.
- Text Fragments – Paragraphs.
- 17 text fragments (v0…v16).
- 17 nodes in Document Graph.

Input parameters for *Document Graph* construction

- ***Parsing*** Delimiters
  - For Plain Text – Newline or Period
  - For HTML – Tags (<p>,<br>,<ul><ol>,<table>… etc.)

- ***Threshold*** for Edge weights
  - Tradeoff of Quality and Performance.
  - Edges with weights lesser, are not added.

- ***Maximum*** Fragment Size
  - Limit on Node Size

Computing edges of Document Graphs

- For every pair of nodes,
  - Common Words are used (stops words – ignored)
  - Thesaurus and stemmer used (rely on Oracle Intermedia Text services)
  - If ***EScore(e)* ≥ *threshold***, an edge is added.

- Special Case
  - Adjacent Text Fragments.
  - Share Close Proximity.
  - Weight = Max (*EScore(e) ,threshold*).

Edge Scoring

- ***EScore***

  A ***tf*idf*** adaptation.

  – *Query Independent*.

  – Edge *e(u,v)*

$$EScore(e) = \frac{\sum_{w \in (t(u) \cap t(v))} \left((tf(t(u), w) + tf(t(v), w)) \cdot idf(w) \right)}{size(t(u)) + size(t(v))}$$

w – common word,

t (*v*) – text fragment corresponding to node *v*.

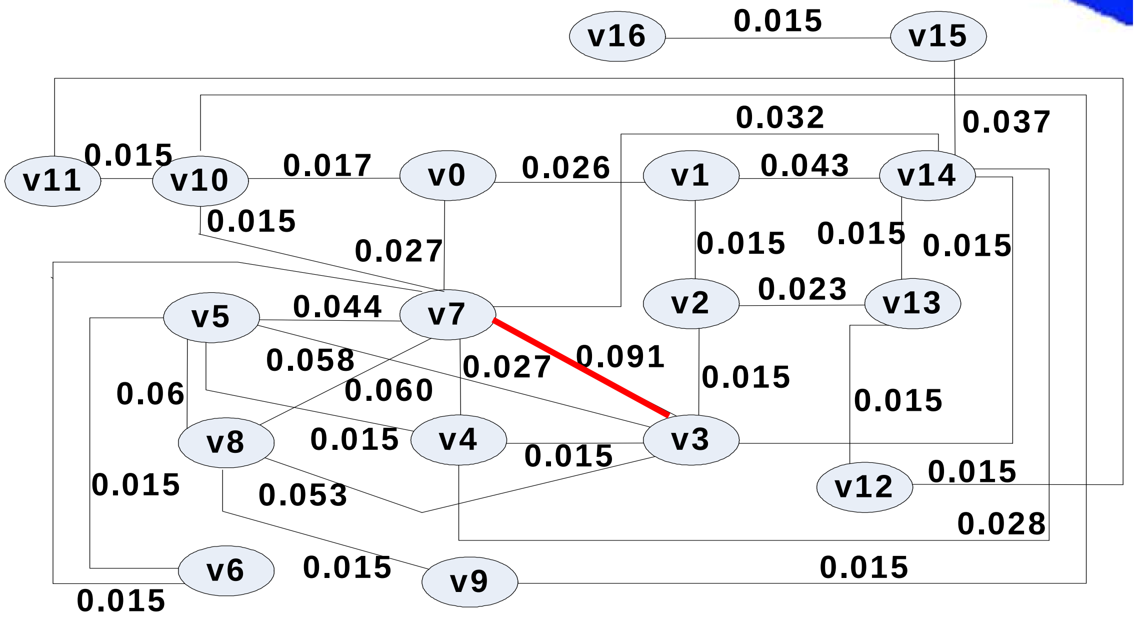
Size (*v*) –number of words in text fragment *t(v)*.

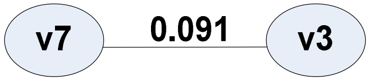
Example (cont'd)

Sample Document

(v0) Brain chip offers hope for paralyzed
(v1) A team of neuroscientists have successfully implanted a **chip** into the **brain** of a quadriplegic man, allowing him to control a computer.
(v2)....
(v3) The **chip**, called BrainGate, is being developed by Massachusetts-based neurotechnology company Cyberkinetics, following **research** undertaken at Brown University, Rhode Island.
(v4) ....
(v5) BrainGate offers the possibility of hitherto unimaginable levels of independence for the severely disabled.
(v6)....
(v7) John Donoghue, professor of neuroscience at Brown and a co-founder of Cyberkinetics in 2001, said that BrainGate could help paralyzed peopled control wheelchairs and communicate using email and Internet-based phone systems.
(v8)....
(v9) ....
(v10) Donoghue's initial **research**, published in the science journal Nature in 2002, consisted of attaching an implant to a monkey's **brain** that enabled it to play a simple pinball computer game remotely.
(v11).....
(v12) "While these results are preliminary, I am extremely encouraged by what has been achieved to-date," said John Mukand of the Sargent Rehabilitation Center, who oversaw the pilot study.
(v13).....
(v14) .....
(v15).....
(v16)......

Document Graph





Common words:

- ***BrainGate***,
- ***Cyberkinetics***

Reasons for high weight

- Rare Words (*idf* is large).

Computing Query-Specific Summaries

- Given a Query, Q and a Document Graph, G:

Summary → Minimal Total Spanning Tree.

Minimal Total Spanning Tree

- *Total* – Every keyword in at least one node (*AND* semantics)
- *Minimal* – To avoid redundancy (Eliminating useless leaves)

Summarization Problem

Given – Document Graph *G* and a Query *Q*

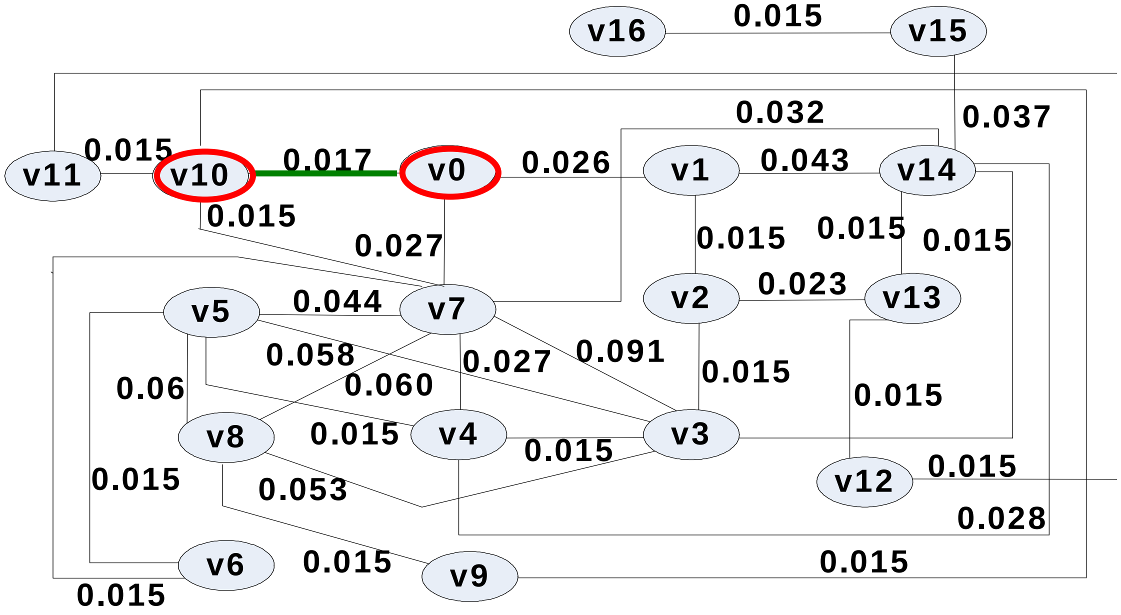
Find – Top (best) Minimal Total Spanning Tree (Summary)

Example

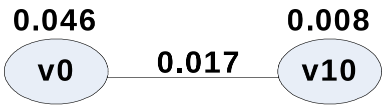
Sample Document

(v0) Brain chip offers hope for paralyzed
(v1) A team of neuroscientists have successfully implanted a **chip** into the **brain** of a quadriplegic man, allowing him to control a computer.
(v2) …
(v3) …
(v4) …
(v5) BrainGate offers the possibility of hitherto unimaginable levels of independence for the severely disabled.
(v6) …
(v7) …
(v8) …
(v9) …
(v10) Donoghue's initial **research**, published in the science journal Nature in 2002, consisted of attaching an implant to a monkey's **brain** that enabled it to play a simple pinball computer game remotely.
(v11) The four-millimeter square **chip**, which is placed on the surface of the motor cortex area of the **brain**, contains 100 electrodes each thinner than a hair which detect neural electrical activity. The sensor is then connected to a computer via a small wire attached to a pedestal mounted on the skull.
(v12) …
(v13)…
(v14)…
(v15) "Here we have a **research** participant who is capable of controlling his environment by thought alone -- something we have only found in science fiction so far," said Friehs.
(v16) …

Document Graph



Top Summary for
"***Brain Chip Research***"



Score = **67.74**

Brain chip offers hope for paralyzed.
⌊ Donoghue's initial **research** published in the science journal Nature in 2002 consisted of attaching an implant to a monkey's **brain** that enabled it to play a simple pinball computer game remotely.

Summary Scoring Function

Requirements

Properties of Good Summaries :

- Highly relevant nodes (fragments) ***improve*** Score.
- Loose semantic Links ***degrade*** Score.
- Large spanning trees get a ***degraded*** Score.
- Based on *Query-dependent* & *Query-Independent* factors.

Summary Scoring

- This function *satisfies* these requirements.
- Best Summary has ***minimum*** score

$$\text{Score(T)} = a \sum_{edge\ e \in T} \frac{1}{EScore(e)} + b \frac{1}{\sum_{node\ v \in T} NScore(v)}$$

a and ***b*** are calibrating parameters.

(a=1 & b=0.5)

Summary Node Scoring

- ***Node* Scoring**

  – Widely used *Okapi weighting*.

  – Query Dependent.

  – ***NScore* (*v*)** = $\sum_{t \in Q,d} \ln \frac{N - df + 0.5}{df + 0.5} . \frac{(k_1 + 1)tf}{(k_1(1 - b) + b\frac{dl}{avdl}) + tf} . \frac{(k_3 + 1)qtf}{k_3 + qtf}$

***N* –** *Number of Documents in the collection.*

tf – *Term Frequency .*

df – *Document Frequency.*

***avdl* –** *Average Document Length.*

Roadmap

- Need for query-specific summaries
- Our approach
  - Building a document graph
  - Definition of summary
  - Rank Summaries
- Efficient computation of summaries
- Evaluation of summarization process
  - Quality
  - Performance
- Related Work
- Conclusions

ALGORITHMS

- Adaptations of **BANKS [ICDE02]** Algorithms
- *Input* : Document Graph *G* and Query *Q*
- *Output* : Minimal Total Spanning trees (Summaries)

- ***Enumeration*** Algorithm.
- ***Expanding Search*** Algorithm.

Pre-computation:

– A Full text Index.

– All Pairs shortest paths for each document graph (edge weight of edge e= 1/Escore(e)).

Roadmap

- Need for query-specific summaries
- Our approach
  - Building a document graph
  - Definition of summary
  - Rank Summaries
- Efficient computation of summaries
- Evaluation of summarization process
  - Quality
  - Performance
- Related Work
- Conclusions

User Surveys

- To *evaluate* the *Quality* of Summaries
- **Subjects** : 15 Students from FIU (all levels & various majors).
- Users evaluate summaries based on their ***Quality***.
- **Rating**: 1 (least descriptive) to 5 (most descriptive)

- **Surveys**
  - Comparison with Google & MSN Desktop.
  - Comparison with DUC 2005 datasets.

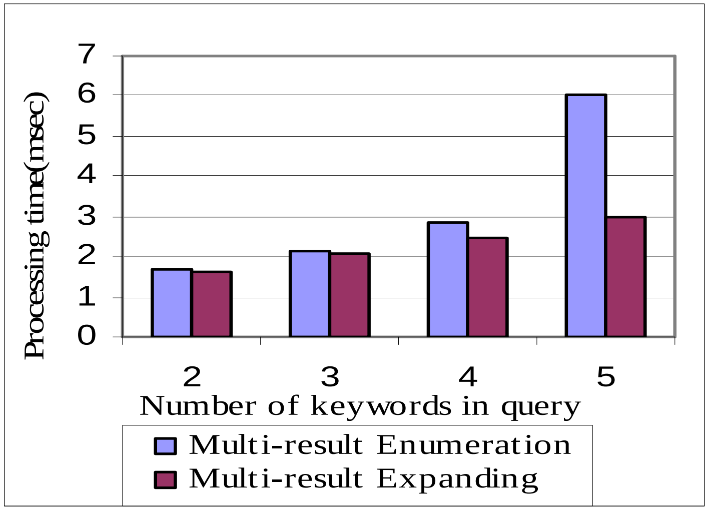
Comparison with Google & MSN Desktop Engines

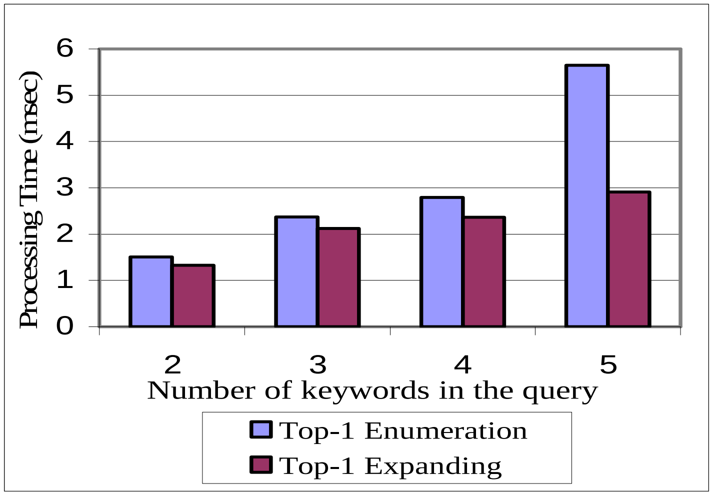
Queries	Google Desktop		MSN Desktop		Our Approach	
	D1	*D2*	*D1*	*D2*	*D1*	*D2*
1	2.33	3.67	2.33	3.67	4.87	3.67
2	2.00	3.33	2.00	3.00	4.33	3.33
3	3.00	2.67	0.67	3.00	4.93	4.00
4	1.67	2.67	1.67	3.00	4.67	4.00
5	2.00	1.67	3.00	1.00	4.00	3.67

Queries	**Document *D1***	**Document *D2***
1	Microsoft worm protection	IT Research awards
2	Anti-virus protection	Algorithms development research
3	Recovering worm deleted files	Software projects
4	Worm affected agencies	Large research grants
5	Deleted computer software	Computer network security project

Performance Experiments

News articles from ***science section*** of ***cnn.com***





Average times to calculate node weights

Number of keywords	2	3	4	5
Time (msec)	5.31	9.37	11.50	17.33

Average ranks of Top-1 Algorithms

Number of keywords	2	3	4	5
Top-1 Enumeration Algorithm.	1.4	1.8	2.1	2.78
Top-1 Expanding Search Algorithm.	1.1	1.3	1.4	1.8

Roadmap

- Need for query-specific summaries
- Our approach
  - Building a document graph
  - Definition of summary
  - Rank Summaries
- Efficient computation of summaries
- Evaluation of summarization process
  - Quality
  - Performance
- Related Work
- Conclusions

Related Work

Document Summarization

- Mostly Query-Independent
- Summarizing Web Pages
  - Berger et.al [SIGIR 2000] synthesizes summaries.
  - Paris et.al [CIKM 2000] uses anchor text (ignores content).
- Splitting Web pages in to blocks
  - Song et.al [WWW2004] Block importance models (learning algorithms)
  - Cai et.al [SIGIR 2004] Block level link analysis
- Document modeled as Graphs
  - Lexrank : Sentence Centrality using link analysis.
  - TextRank: “representative” sentences using link analysis.

Keyword Search in Data Graphs

- BANKS [ICDE 2002]: group-steiner tree problem
- DISCOVER, DBXplorer.
- XRANK[2003]: search in XML documents.

Conclusions

- Method for Query-Specific Summarization.
- Exploiting inherent structure of documents for the purpose of Summarization.
- Enhanced User Satisfaction – User Surveys.

A Prototype of the System available at:

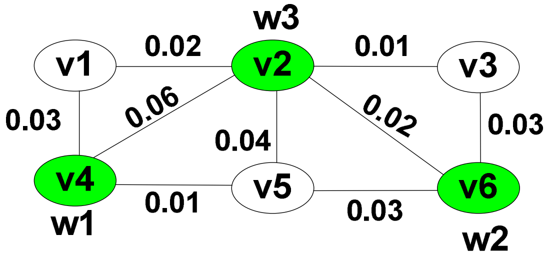
http://dbir.cs.fiu.edu/summarization

Thank You !!!

Questions ???

Enumeration Algorithm

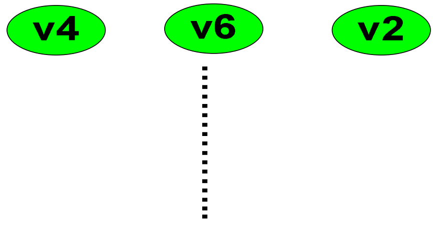
SAMPLE DOCUMENT GRAPH



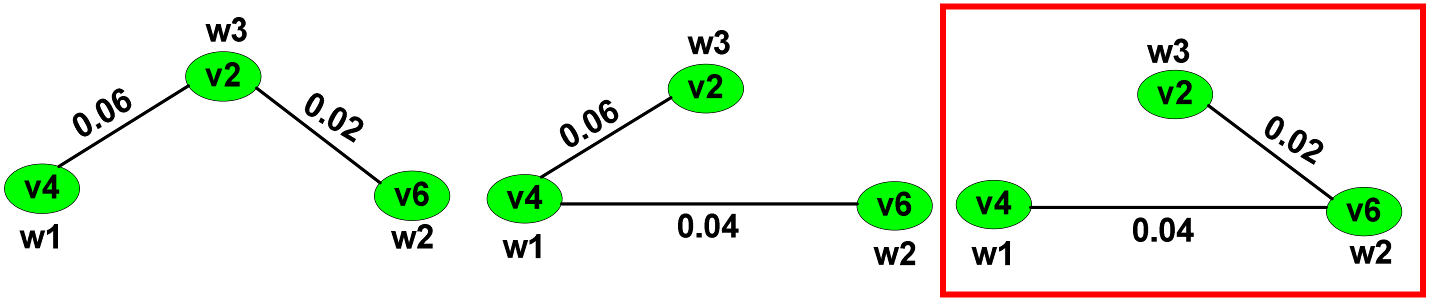
● Keyword Node

Q = {w1,w2,w3}

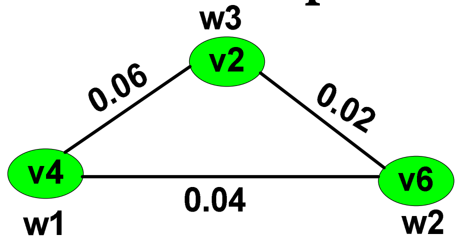
Minimal Node Combinations



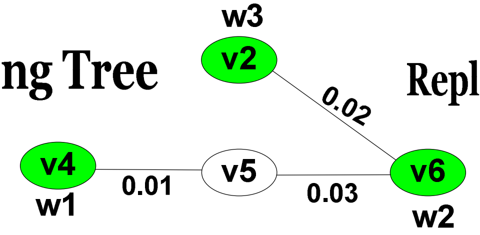
Possible Spanning Trees



Closure Graph



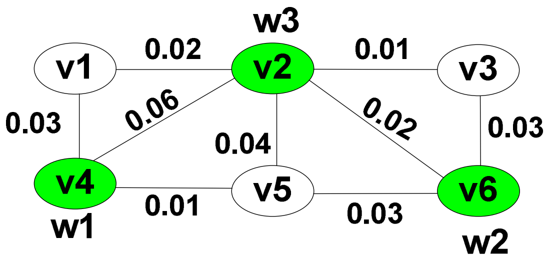
Best Spanning Tree



Replacing edges with shortest paths

Expanding Search Algorithm

SAMPLE DOCUMENT GRAPH



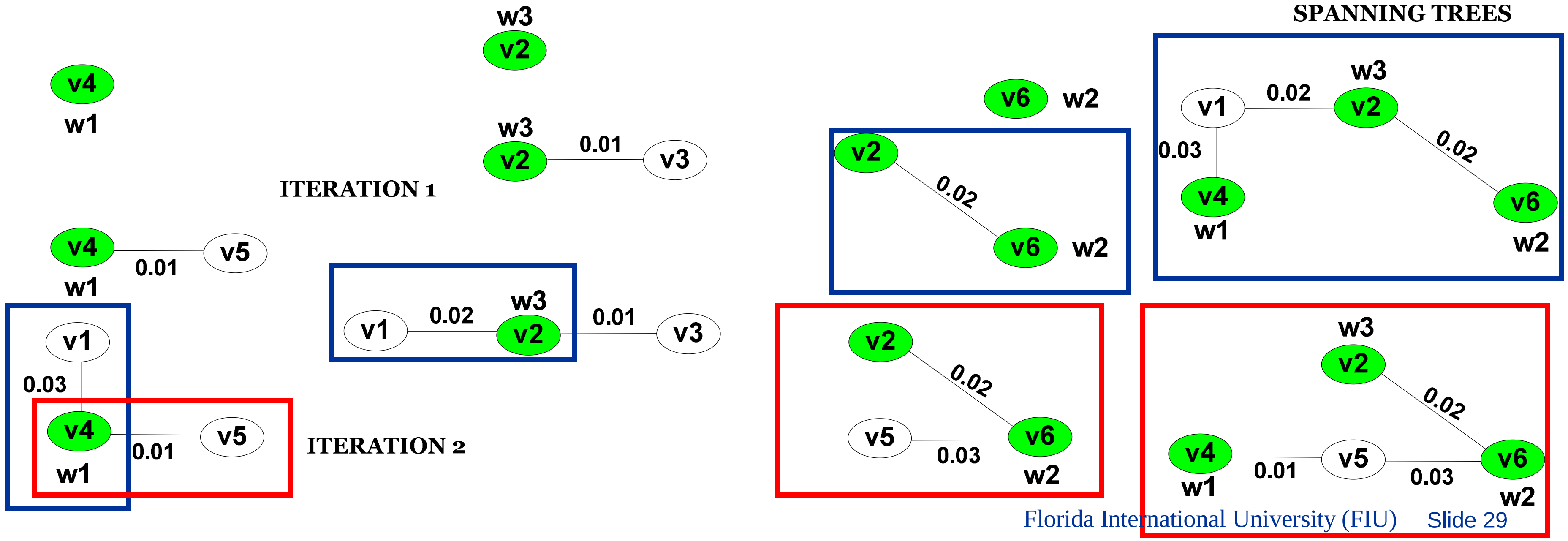
● Keyword Node

Q = {w1,w2,w3}

EXPANDING AREA
of keyword node v4

EXPANDING AREA
of keyword node v2

EXPANDING AREA
of keyword node v6



SPANNING TREES

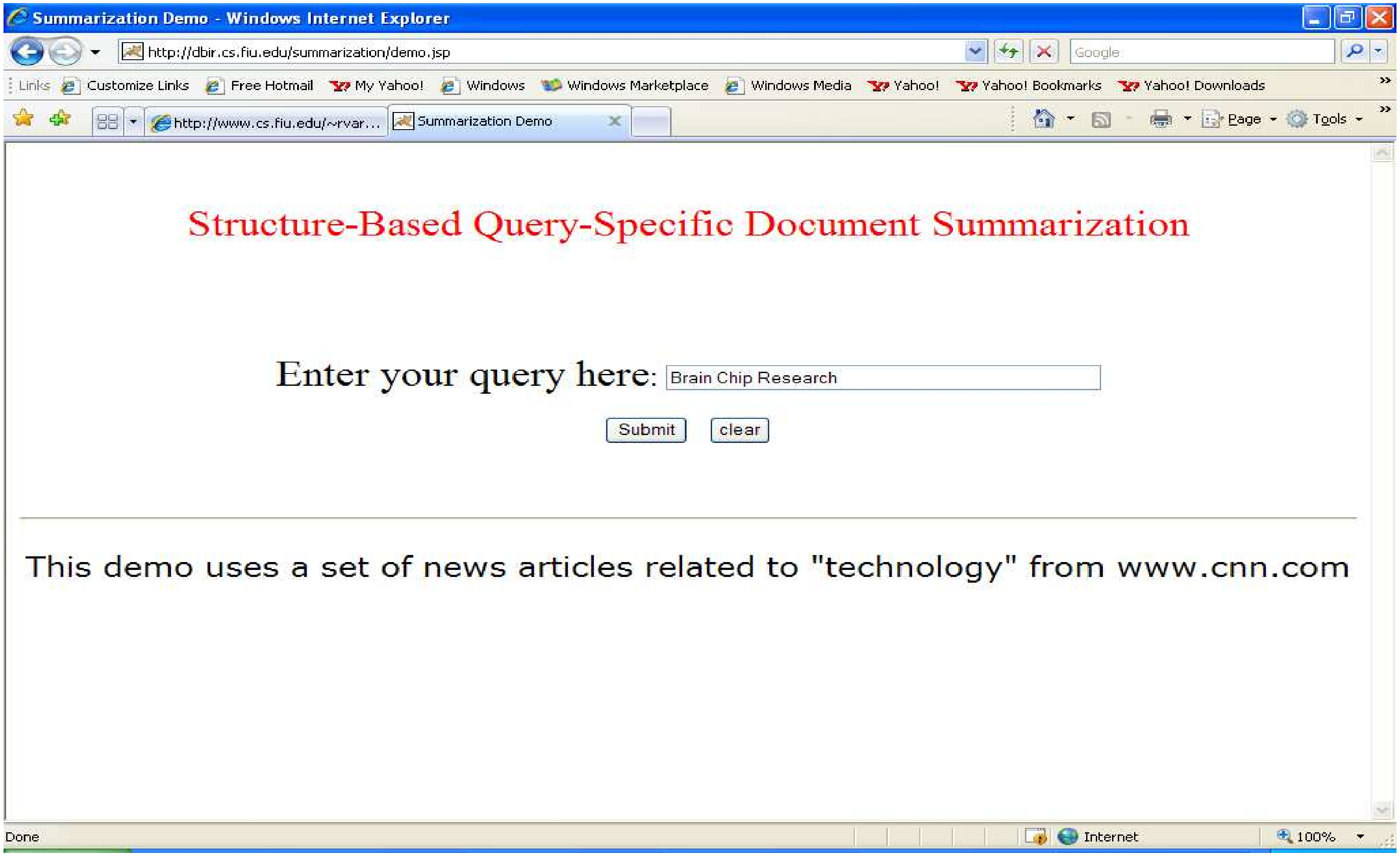
ITERATION 1

ITERATION 2

Comparison with DUC peers

Query 1 (*International Organized Crime*) DUC Topic ID: d301i			Query 2 (*Women in Parliaments*) DUC Topic ID: d321f		
Doc. ID	DUC Peer	Our approach	Doc. ID	DUC Peer	Our Approach
FT941-3237	2.33	4.66	FT921-7786	4.00	2.50
FT944-8297	2.50	3.33	FT922-190	2.00	4.00
FT931-3563	2.83	3.00	FT921-937	2.00	4.33
FT943-16477	4.00	4.17	FT922-13353	2.83	4.17
FT943-16238	3.67	3.67	FT921-74	2.33	3.67

DEMO



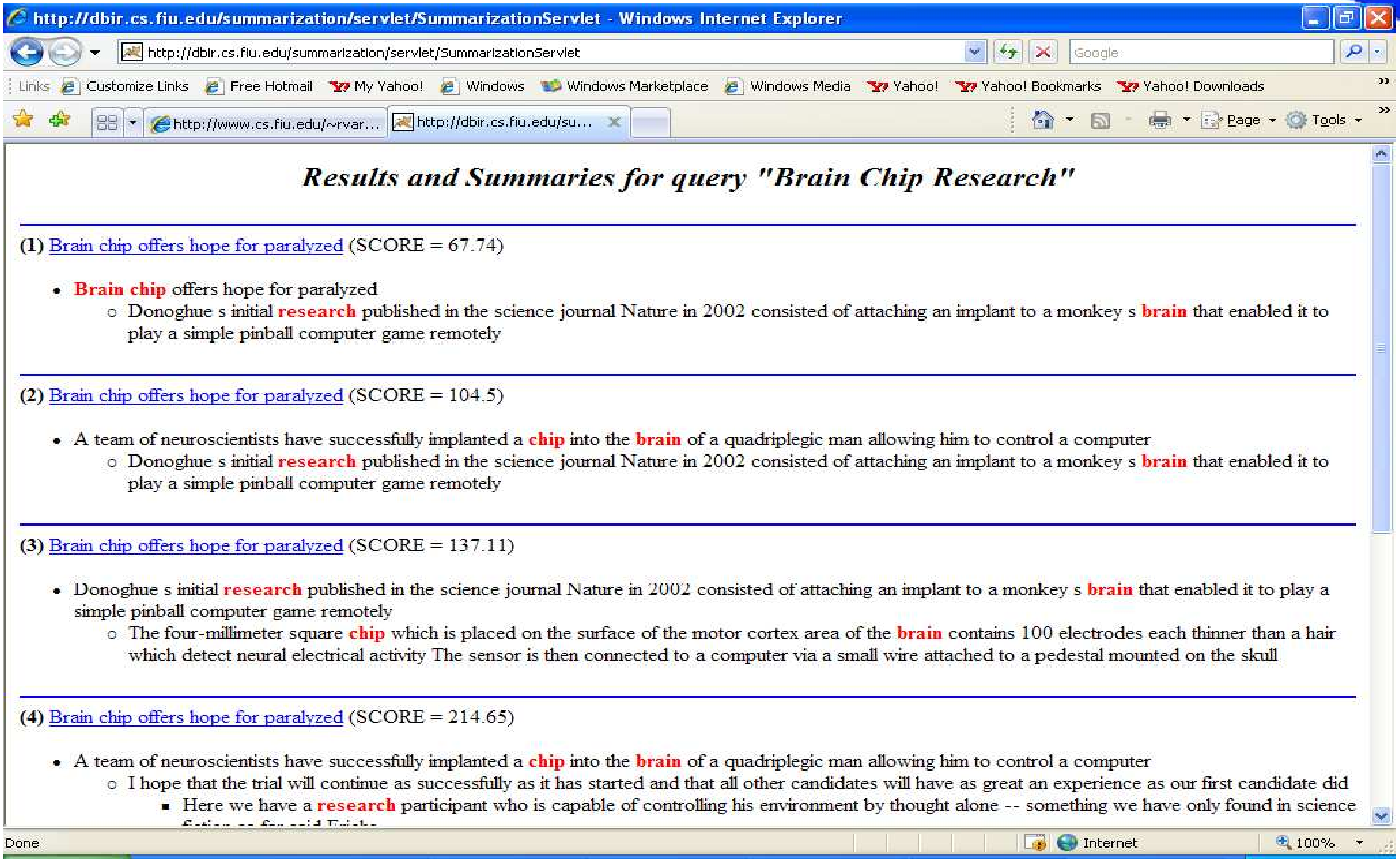
Structure-Based Query-Specific Document Summarization

Enter your query here: Brain Chip Research

Submit clear

This demo uses a set of news articles related to "technology" from www.cnn.com

DEMO


http://dbir.cs.fiu.edu/summarization/servlet/SummarizationServlet - Windows Internet Explorer
http://dbir.cs.fiu.edu/summarization/servlet/SummarizationServlet
Google
Links
Customize Links
Free Hotmail
My Yahoo!
Windows
Windows Marketplace
Windows Media
Yahoo!
Yahoo! Bookmarks
Yahoo! Downloads
http://www.cs.fiu.edu/~rvar...
http://dbir.cs.fiu.edu/su...
Page
Tools
Done
Internet
100%

Results and Summaries for query "Brain Chip Research"

(1) Brain chip offers hope for paralyzed (SCORE = 67.74)

- **Brain chip** offers hope for paralyzed
  - Donoghue s initial **research** published in the science journal Nature in 2002 consisted of attaching an implant to a monkey s **brain** that enabled it to play a simple pinball computer game remotely

(2) Brain chip offers hope for paralyzed (SCORE = 104.5)

- A team of neuroscientists have successfully implanted a **chip** into the **brain** of a quadriplegic man allowing him to control a computer
  - Donoghue s initial **research** published in the science journal Nature in 2002 consisted of attaching an implant to a monkey s **brain** that enabled it to play a simple pinball computer game remotely

(3) Brain chip offers hope for paralyzed (SCORE = 137.11)

- Donoghue s initial **research** published in the science journal Nature in 2002 consisted of attaching an implant to a monkey s **brain** that enabled it to play a simple pinball computer game remotely
  - The four-millimeter square **chip** which is placed on the surface of the motor cortex area of the **brain** contains 100 electrodes each thinner than a hair which detect neural electrical activity The sensor is then connected to a computer via a small wire attached to a pedestal mounted on the skull

(4) Brain chip offers hope for paralyzed (SCORE = 214.65)

- A team of neuroscientists have successfully implanted a **chip** into the **brain** of a quadriplegic man allowing him to control a computer
  - I hope that the trial will continue as successfully as it has started and that all other candidates will have as great an experience as our first candidate did
    - Here we have a **research** participant who is capable of controlling his environment by thought alone -- something we have only found in science fiction so far said Friehs