WEBSITE TAG PROPAGATION

AMULYA K – MT2012017

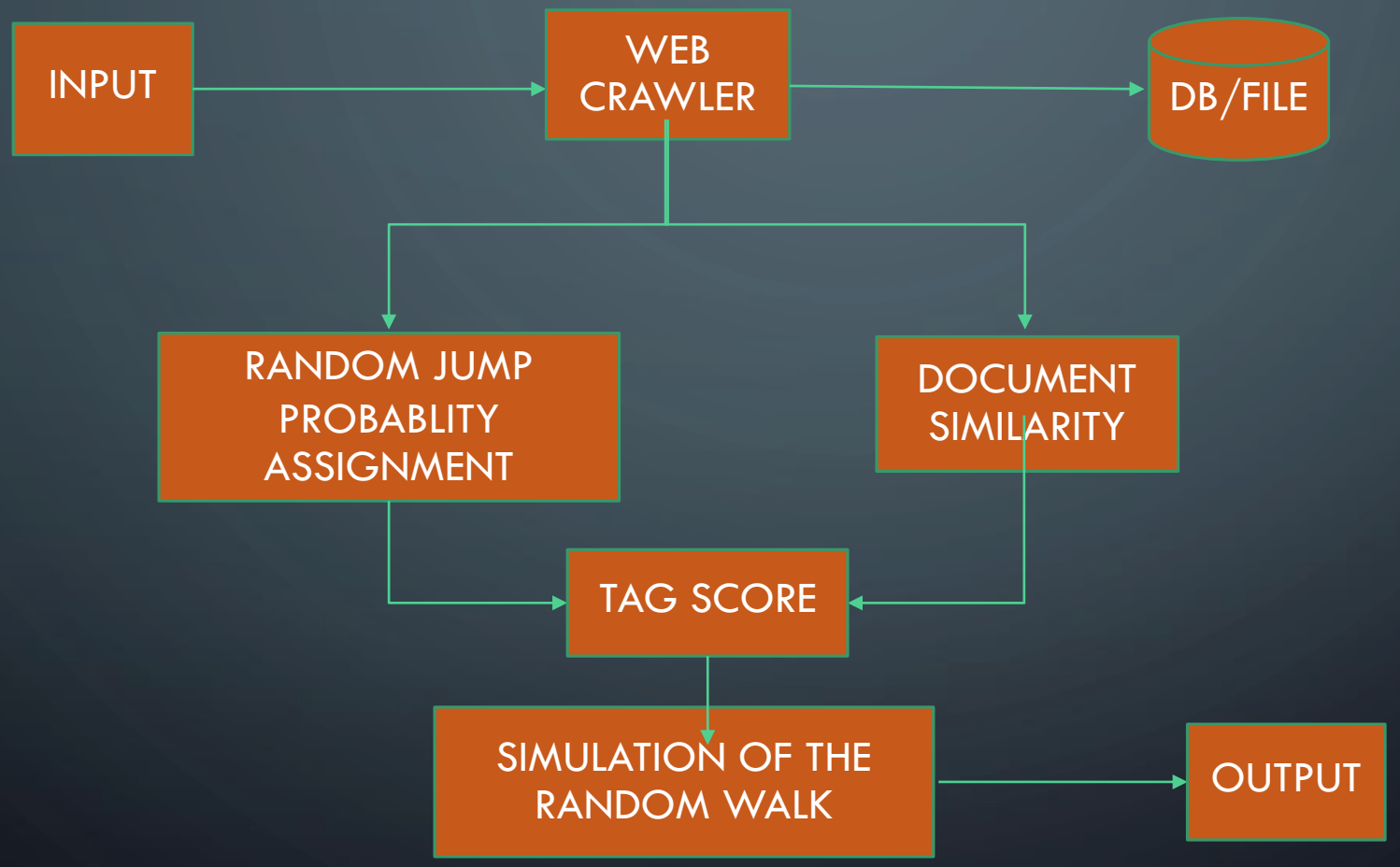
N PUNEETH – MT2012083

SINDHU PRIYADARSHINI – MT2012134

OBJECTIVE

- To accept a user given website and the associated tags
- To create a web graph starting from that initial website to a certain number of hops.
- To Propagate the tags within that web graph.
- To Use Link Analysis and any other relevant technique to create the system.

SYSTEM ARCHITECTURE



RANDOM SURFER MODEL

$$Pij = \frac{Xij(1-\alpha)}{\sum_j Xij} + \frac{\alpha}{N}$$

Let us have a web graph represented as graph G with N nodes, (1.....N).
We have an Adjacency Matrix: A
Transition Probability Matrix: P
Let α be the probability of teleport operations.
If a row of A has no 1s, set each element to 1/N.

DOCUMENT SIMILARITY

- Cosine Similarity Metric

Document a:

```
A B C A A B C. D D E A B. D A B C B A.
```

Document b:

```
A B C A A B C. D A B C B A.
```

Vector a:

```
A:6, B:5, C:3, D:3, E:1
```

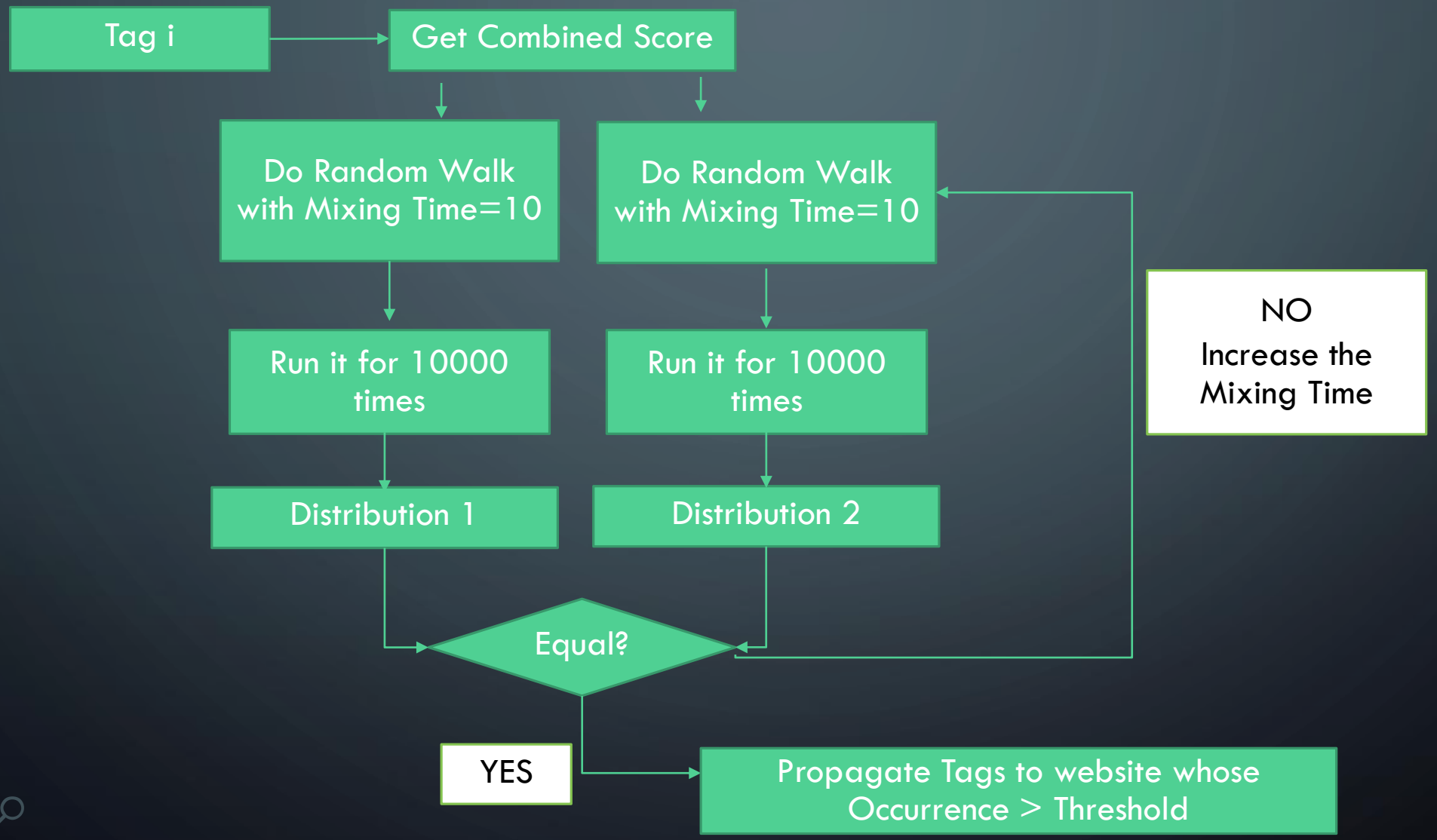
Vector b:

```
A:5, B:4, C:3, D:1, E:0
```

Which result in the following similarity measure:

```
(6*5+5*4+3*3+3*1+1*0)/(Sqrt(6^2+5^2+3^2+3^2+1^2)
Sqrt(5^2+4^2+3^2+1^2+0^2))= 62/(8.94427*7.14143)   = 0.970648
```

ALGORITHM


Tag i
Get Combined Score
Do Random Walk with Mixing Time=10
Do Random Walk with Mixing Time=10
Run it for 10000 times
Run it for 10000 times
Distribution 1
Distribution 2
Equal?
NO
Increase the Mixing Time
YES
Propagate Tags to website whose Occurrence > Threshold

PARAMETERS

- Number of Samples
- Standard Distribution
- Threshold Value
- Mixing Time

ISSUES

- URL errors
  - Malformed URL
  - Images, Audio, Videos, documents etc.
  - Social Sites
  - Redirected URLs
- Heap Size

TESTING STRATEGY

- Every Output of the system was manually checked.
- Each correct website-tag pair was annotated green.
- Some of the website tag pairs were ambiguous, so they were annotated as yellow.
- Some of the website-tag pairs were wrong, they were annotated with color red.

RESULTS

- The websites that were recommended by the system were relevant, primarily because of the two metrics which we used in our system.
- The lack of semantic understanding was the reason for the errors which we encountered.
- 89.45% accuracy.

	Total Output	Incorrect	Ambiguous	Correct	Correct %
Steven Gerrard	244	9	18	217	88%
Cricinfo	10		2	8	80
The Hindu	7			7	100
John Lennon	4			4	100
Uttarahalli	190	4	15	171	90%
	455	13	35	407	89.45

Correct – 89.45% (407/455) Incorrect – 2.8% (13/455)

CONCLUSION

- Simple probabilistic graphical models can be highly efficient. Random Surfer Model.
- The Propagation of the tags can be done with limited information.
- The User given tags might not always be right.
- Document Similarity/some other page level metric is vital.
- Web is a dynamic entity.

.

FUTURE WORK

- Google Uses more than 200 signals to rank their webpages.
- More of these metrics could be used to improve upon the system.
- Domain Name, Title Tags, Meta-tags etc.
- Machine Learning Based
- Parallelization, Distributed system