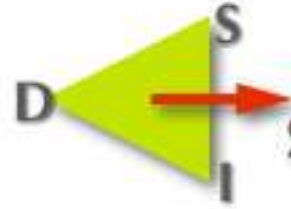




Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

Classifier

Travail effectuée

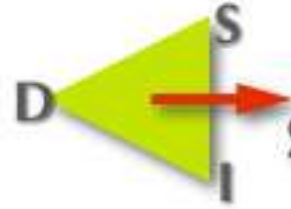
Conclusion

# Le Projet Ronin

Mansart Jean-Sébastien



Università degli Studi di Firenze



**Dipartimento di  
Sistemi e Informatica**

Presentation

Theme du projet

Classifier

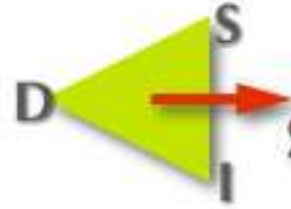
Travail effectue

Conclusion

# Presentation



Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

→ Laboratoire

Universite de Florence

Laboratoire “Machine Learning & Neural Networks  
Group”

Professeurs :

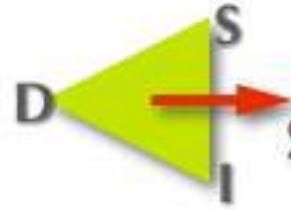
Paolo Frasconi, Simone Marinai, Giovanni Soda.

Post Doctorat :

Fabrizio Costa, Enrico Francesconi.



Università degli Studi di Firenze



**Dipartimento di  
Sistemi e Informatica**

Presentation

Theme du projet

Classifier

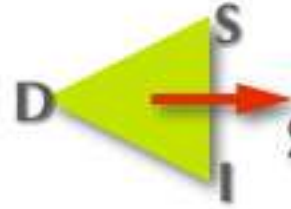
Travail effectue

Conclusion

# Theme du projet



Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

→ Presentation

generale

Focused

crawling

Architecture

Differences

Avantages

Membres du projet:

Paolo Frasconi, Fabrizio Costa, Jimmy Dubuisson

Developpement d'un moteur de recherche distribue.

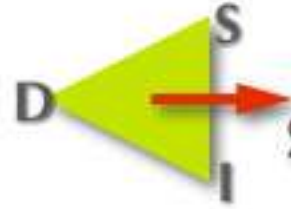
La distribution d'un tel systeme pose des problemes de coordination.

Solutions:

- P-Grid
- Focused crawling



Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

Presentation  
generale

→ Focused  
crawling

Architecture

Differences

Avantages

Permet de focaliser la recherche sur un seul topic.

Introduction du concept de web communautaire.

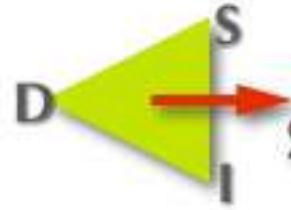
Le web vu comme un graph.

Metrique definie:

- Relation directe: une page web a un lien vers une autre page
- Co-citation: une page lie deux autres pages
- Social filtering: une page est lie par deux autres pages



Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

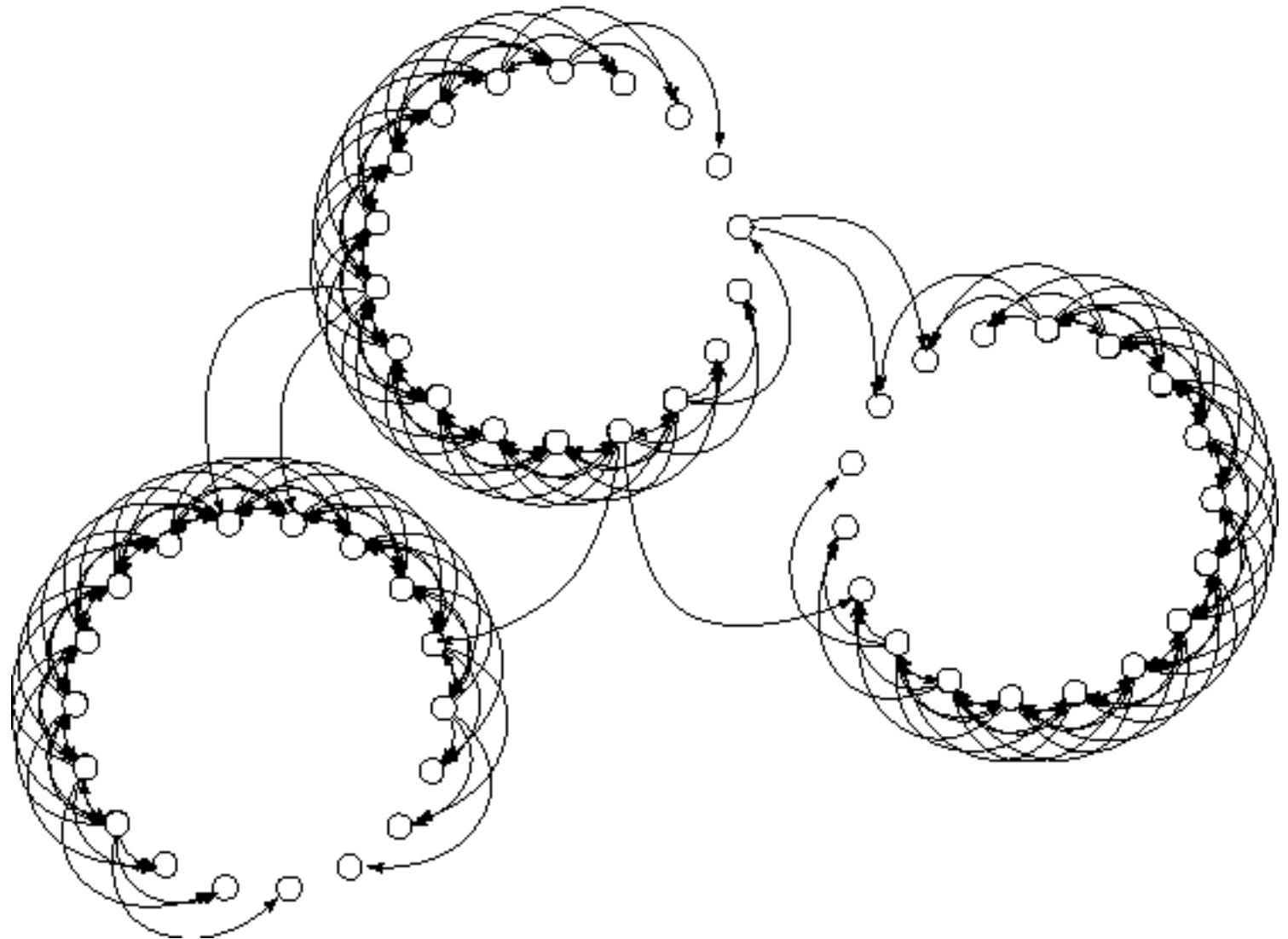
Presentation  
generale

→ Focused  
crawling

Architecture

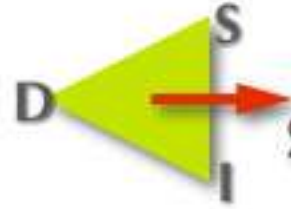
Differences

Avantages





Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

Presentation  
generale

Focused  
crawling

→ Architecture

Differences

Avantages

Croissance rapide du web

Les architectures centralises ne pourront plus suivre.

Solution:

Utiliser un systeme distribue: Grid/web-services

Probleme:

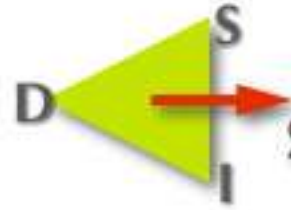
Coordination : Focused crawling.

Grid:

Permet de partager des donnees, de l'espace disque, de la memoire et les connections reseaux a travers des organisation heterogene et dispersees geographiquement.



Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

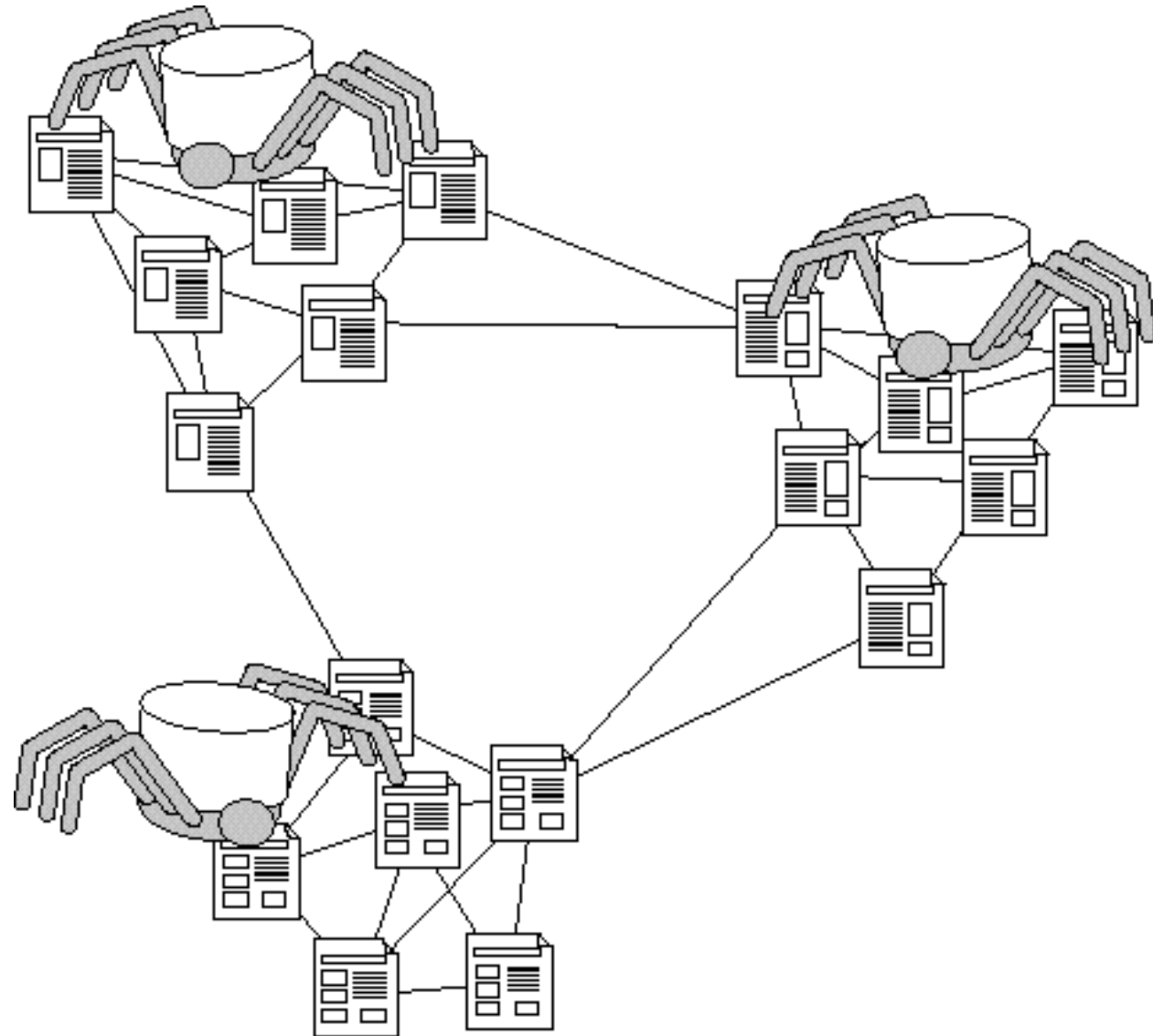
Presentation  
generale

Focused  
crawling

→ Architecture

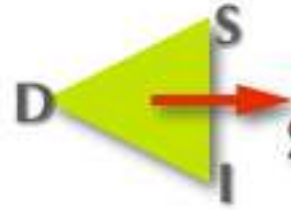
Differences

Avantages





Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

Presentation  
generale

Focused  
crawling

→ Architecture

Differences

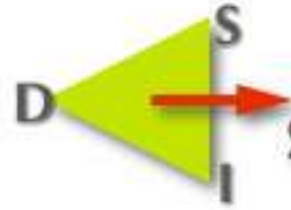
Avantages

Trois niveaux:

- Crawler  
explorent le web et determinent quels document recuperer.
- Proxy  
recois les documents du fetcher et les donne aux crawlers.
- Fetcher  
recuperent les pages web, et renvoient les documents au proxy.



Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

Presentation  
generale

Focused  
crawling

Architecture

→ Differences

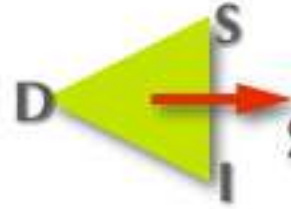
Avantages

Architecture centralise (Google):

- Probleme de ressources et de temps.
- Mise a jour des informations.
- Recupere toutes les informations pour repondre a toutes les requetes possibles.
- Enregistre toutes les pages web dans une base de donnees.



Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

Presentation  
generale

Focused  
crawling

Architecture

Differences

→ **Avantages**

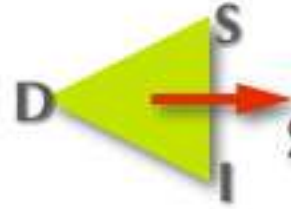
Pas de limitation theorique pour les ressources.

Plus il y a d'ordinateurs connecte, plus la  
couverture du web est importante.

Possibilite d'avoir un moteur de recherche specialise  
dans un seul domaine.



Università degli Studi di Firenze



**Dipartimento di  
Sistemi e Informatica**

Presentation

Theme du projet

Classifier

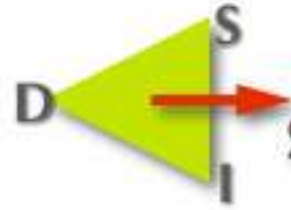
Travail effectue

Conclusion

# Classifier



Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

Classifier

→ Presentation

But du  
classifier

Training-set

Module qui m'a été assigné.

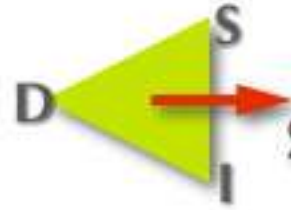
Utilisation du focused crawling

Question: comment évaluer la précision avec laquelle le focused crawler récupère les pages web?

Faire un autre classificateur basé sur une autre technologie.



Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

Classifier

Presentation

→ But du

classifier

Training-set

Ce classifieur ne sera pas intégré dans le projet Ronin.

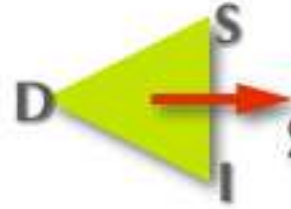
Sert uniquement pour tester l'algorithme du focused crawling, afin de l'améliorer si besoin est.

Classifier basé sur un algorithme de machine learning:

Définir un training-set pour apprendre à l'algorithme à reconnaître un exemple positif d'un exemple négatif.



Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

Classifier

Presentation

But du  
classifier

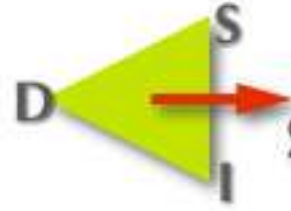
⇒ Training-set

L'elaboration du training-set s'est faite en plusieurs etapes:

- Definition d'une liste de mots cles
- Recuperation de page web a partir de la liste de mots-cles
- Trouver un moyen sur pour determiner si une page est un exemple positif ou non.
- Entrainer l'algorithme.



Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

Classifier

Presentation

But du

classifier

⇒ Training-set

Definition des mots clés

Recuperation d'une bibliotheque de bibtex au format  
texte.

Extraction des titres des articles

Suppression des “stops words”

Creation de couples de mots-cles

“Toward the evolution of dynamical neural networks for minimally cognitive  
behavior”

“toward evolution dynamical neural networks minimally cognitive behavior”

“toward evolution”

“evolution dynamical”

“dynamical neural”

“neural networks”

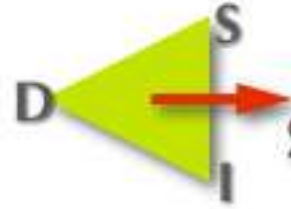
“networks minimally”

“minimally cognitive”

“cognitive behavior”



Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

Classifier

Presentation

But du  
classifier

⇒ Training-set

Recuperation des pages web.

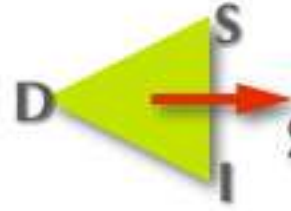
Recuperer des pages web contenant au moins un lien vers un fichier pdf.

Utilisation de l'API google: web-service.

- Constitution d'une liste contenant l'URL de fichier pdf
- Recherche des pages web qui ont un lien vers l'URL du pdf.



Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

Classifier

Presentation

But du

classifier

→ Training-set

Positif ou negatif ?

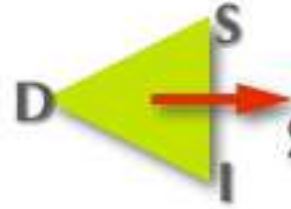
Liste de pages web contenant au minimum un lien vers un fichier pdf.

Deux facons:

- Rechercher dans la page web certains criteres (titre d'un article, nom de conference, annee, nom du journal)
- Examiner le fichier pdf: si il y a un abstract et une table des references.



Università degli Studi di Firenze



**Dipartimento di  
Sistemi e Informatica**

Presentation

Theme du projet

Classifier

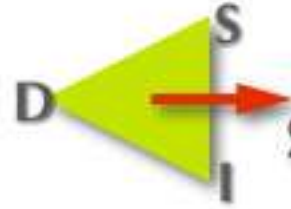
Travail effectue

Conclusion

# Travail effectue



Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

Classifier

Travail effectuée

→ Methodes de  
travail

Systeme de  
communication

Technologie

Data-Set

Communication principalement MSN messenger,  
email, video-conference (plus rare).

Utilisation d'un CVS pour le projet

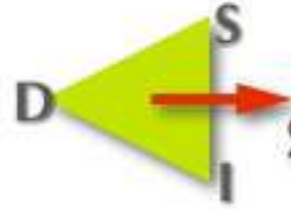
Documentation:

Production de documentation pour l'installation  
de logiciels, exemples d'utilisations, exemples de  
programmes.

Generation de documentation HTML avec  
Dioxygen et Javadoc.



Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

Classifier

Travail effectuée

Methodes de  
travail

→ Systeme de  
communication

Technologie

Data-Set

Mise en commun d'articles scientifiques:

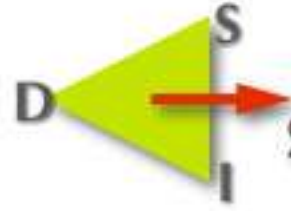
- site web ou déposer des articles, ajouter des commentaires sur les articles.
- definition de groupes: un groupe = un theme de recherche
- identification des utilisateurs
- utilisation de Plone

Video conference:

Installation d'un logiciel (VRVS) pour faire de la visio conference avec les personnes exterieurs.



Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

Classifier

Travail effectue

Methodes de  
travail

Systeme de  
communication

→ Technologie

Data-Set

## Recherche de librairie C++

- Trouver une librairie qui permet de se connecter a une URL et d'effectuer des actions dessus.
- Trouver une autre librairie gerant les expressions regulieres

## Web-services

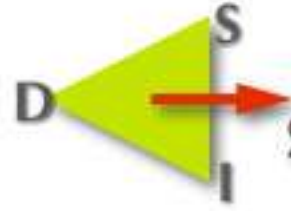
Autre moyen de distribuer le moteur de recherche.

Crawler = client

Fetcher = serveur



Università degli Studi di Firenze



Dipartimento di  
Sistemi e Informatica

Presentation

Theme du projet

Classifier

Travail effectuée

Methodes de  
travail

Systeme de  
communication

Technologie

→ Data-Set

Utilise pour les tests du moteur de recherche.

Data-set compose de pages web

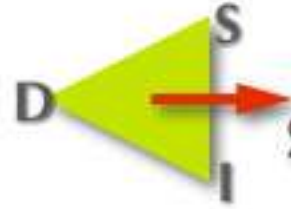
Evite les acces a Internet avec tous les problemes que cela impose: charge des serveur, lenteur, indisponibilite, etc...

1600000 documents

Extraction de certaines donnees et stockage dans une base de donnees.



Università degli Studi di Firenze



**Dipartimento di  
Sistemi e Informatica**

Presentation

Theme du projet

Classifier

Travail effectue

**Conclusion**

# Conclusion

