**각종 매칭 어플리케이션에 사용된 알고리즘 및 원리**

**소프트웨어 융합 창의설계 5주차 자료 및 정보 수집**

정보 수집 방법: 문헌 및 인터넷 조사

**매칭 알고리즘**

매칭 알고리즘은 이용자의 정보를 분석해 저장한 후, 수집된 추가정보를 학습시켜 보다 빠르고 효과적인 정보 간 매칭 시스템을 제공한다. 대표적으로는 게일-섀플리 알고리즘(Gale-Shapley Algorithm)과 아호 코라식 알고리즘(Aho-Corasick string matching algorithm) 이 있다.

**1. 게일-섀플리 알고리즘(Gale-Shapley Algorithm)**

**: 1대1의 추천 알고리즘**

**1) Background: Stable Marriage Problem(또는 Stable Matching Problem, 이하 SMP)**

동등한 인원 수가 있는 두 집단이 있을 때, (예를 들자면 n명의 남성과 n명의 여성, 혹은 n명의 학생과 n개의 대학) 한 집단의 participant는 다른 집단의 participants을 선호도에 따라 각자 순위를 매긴다. 한 집단의 participant는 반드시 다른 집단의 participant와 매칭되어야 한다.

이 경우 matching은 안정적이지 않다

① 첫 번째 set의 participant인 A는 두 번째 set에 속한 자신의 짝보다 B를 더 선호한다.

② B도 마찬가지로 첫 번째 set에 속한 자신의 짝보다 A를 더 선호한다.

즉, stable matching 은 두 개의 set에서 생성된 모든 match들 중, 각자의 짝보다 다른 participant를 더 선호하는 순서쌍 (A, B) 가 하나도 존재하지 않을 때 성립한다. 이 stable matching은 언제나 존재하며, 본 알고리즘은 stable한 matching을 찾아내어 문제를 해결하는 알고리즘이다.

**2) Solution**

이 알고리즘은 여러 round(혹은 iteration)을 거친다.

1st round

a)각 남성은 자신이 가장 좋아하는 여성에게 청혼한다.

b)1명에게 프로포즈를 받은 여성은 그 남성과 잠정적으로 매칭이 되고, 2명 이상에게 프로포즈 받은 여성은 그 중에서 자신이 가장 선호하는 사람을 골라 잠정적으로 매칭이 ​된다.  이제 잠정적으로 매칭된 커플, 그리고 매칭이 되지 않은 남성과 여성이 남는다.

2nd round

a) 1단계에서 잠정적으로 매칭되지 않은 남성은 자신이 지금까지 청혼하지 않았던 ​여성들 중에서 가장 선호하는 사람에게 청혼한다.

b) 각 여성은 잠정적으로 매칭되었던 남성과 새롭게  프로포즈한 남성 중에서 가장 선호하는 사람을 골라 다시 잠정적으로 매칭이 된다. 모든 남성과 여성들이 잠정적으로 매칭이 되면, 그것이 매칭 결과가 되며 알고리즘이 ​끝난다.

결론적으로, 알고리즘이 끝난 후 짝이 없는 남녀는 없으며, 모든 남성들은 누군가에게는 청혼을 했을 것이고, 모든 여성은 누군가에게는 청혼을 받았을 것이다. [[1]](#endnote-1)

**3) Software package 에서의 사용**

Gale-Shapley Algorithm은 R, Python, MATLAB 등의 프로그래밍 언어와 API 등의 Matching tools의 library 나 package 등에 포함되어 있다.

**4) 활용 사례**

본 알고리즘은 지속적으로 업데이트되고 발전되며 의대생과 병원을 이어주는 레지던트 지원 프로그램, 뉴욕 공립학교 배정 프로그램, 장기기증 매칭 시스템, 데이팅 앱 매칭 프로그램 등에서 활발히 사용되고 있다.

틴더(Tinder) 등의 데이팅 앱에서는 사용자가 이성에게 받은 Likes의 수와 사용자가 이성에게 보낸 Likes를 기반으로 데이터를 구축해 각 사용자들의 우선순위와 선호도를 파악한 뒤 알고리즘을 적용한다. [[2]](#endnote-2)

**2. 아호 코라식 알고리즘(Aho-Corasick string matching algorithm)**

**: 1대 다의 문자열 매칭 알고리즘**

**1) Backgrounds: String Matching Problem**

String matching problem은 더 큰 문자열에서 주어진 string pattern(문자열 패턴)을 찾는 문제이다. 이 string matching problem은 컴퓨터 과학 분야에서 가장 기본적인 주제이고, 많은 어플리케이션들의 제작에서 활용되고 있다. 그 중 가장 유명한 Multipattern String Matching Algorithm은 “Aho-Corasick”이라는 Alfred V. Aho와 Margaret J. Corasick이 고안한 문자열 검색 알고리즘(매칭 알고리즘)이다.

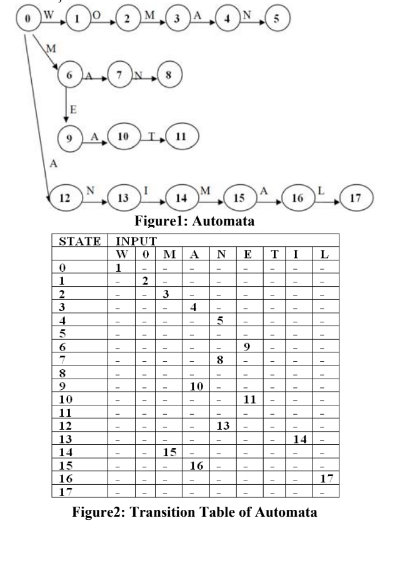
**2) Solution: Aho-Corasick Example**

**Automata**

for Patterns Set= {WOMAN, MAN, MEAT, ANIMAL}

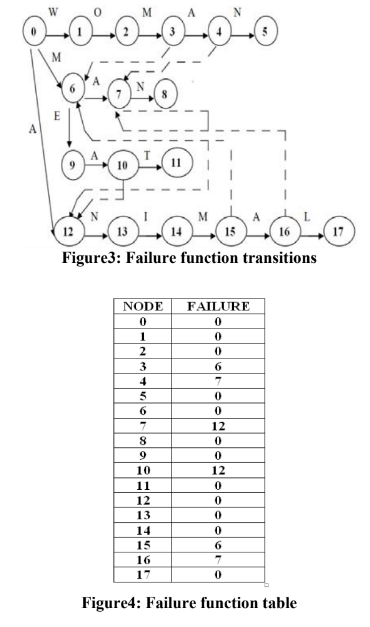
우리의 목적은 4개의 문자열로 이루어진 이 set에서 **주어진 문자열(예시: WOMANETIMEAT)**과 공통된 패턴을 찾는 것이다. 우선 위 pattern set를 트리 구조를 이용해 Automata 로 구성하자.

동그라미는 ‘노드’를 나타내고, 화살표 위의 값은 push, 즉 입력이다. 예를 들어 노드에서 ‘W’라는 입력을 받으면 노드 1로 이동하는 것이다.



**Failure Function**

아래 표에서 Failure는 매칭 실패 시 다음으로 가서 시도해야 할 노드를 의미한다.



예시를 들어보자. 현재 우리는 주어진 문자열 ANIMAN내에서 pattern set의 4개 문자열들과 매칭되는 패턴이 있는지 찾아볼 것이다.

우선 **루트 노드 0에서 시작**한다. 노드 0->12->13->14->15->16까지 차례대로 따라간다. 여태까지 A,N,I,M,A 이 차례대로 매칭되었다.

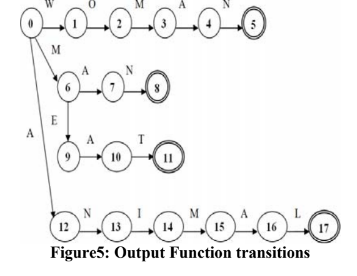
그러나 노트 16에서 17로 넘어갈 수 없다. 노드 17로 넘어가려면 문자 ‘L’이 입력되어야 하는데, 우리한테 남은 문자는 ‘N’이기 때문이다.

Failure function table을 보자. 노드 16에서 failure 값은 7이다. 따라서 노드 16에서 노드 7로 넘어간 뒤, 우리에게 남은 문자 ‘N’을 입력받아 노드 7에서 노드 8로 넘어간다.

ANIMAN의 모든 문자를 살펴봤고, 우리는 **‘ANIMAN’에서 pattern set의 문자열 ‘MAN’을 발견할 수 있음**을 확인할 수 있다.

**Output Function**

마지막 노드에 도착했을 때 어떤 문자열을 발견하게 되는지 알려주는 함수이다.

테이블이(가) 표시된 사진

자동 생성된 설명

**Aho-Corasick Searching Phase**

**주어진 문자열이 WOMANETIMEAT 일 때, 어떤 패턴 문자열이 존재하는지 찾아보자.**

테이블이(가) 표시된 사진

자동 생성된 설명

노드 0->1->2->3->4->5로 차례대로 이동, 노드 5에 도착하여 output function을 통해 문자열 WOMAN 과 MAN을 발견했음을 알 수 있다.

다시 시작 노드 0으로 이동한다. E 로는 이동할 수 있는 노드가 없다. 노드 0의 failure function은 0이므로 노드 0으로 다시 이동하지만 제자리이다. 따라서 문자 E는 매칭에 실패한다. T, I도 마찬가지이다.

그리고 다시 노드 0->6->9->11로 차례대로 이동, 노드11에 도착하여 output function을 통해 문자열 MEAT를 발견했음을 확인할 수 있다.

**따라서 문자열 WOMANETIMEAT에서는 문자열 WOMAN, MAN, MEAT를 발견할 수 있다.**

**3) 활용 사례**

**Intrusion Detection**

침입 탐지 시스템(Intrusion Detection System, IDS)은 일반적으로 시스템에 대한 원치 않는 조작을 탐지하여 준다. 이 시스템은 악의적인 것으로 추정되는 트래픽 또는 애플리케이션 데이터의 패턴을 감시하여 침입을 식별한다.

**Detecting Plagiarism**

아호-코라식 알고리즘을 이용하여 복사/붙여넣기, 짜깁기, 유사한 표현, 전체적인 구조의 유사도 등을 탐색해 표절 여부를 판단할 수 있다.

**Bioinformatics**

본 알고리즘은 바이오인포매틱스 분야에서 생물학적 데이터(염기서열, DNA/RNA 등)를 찾아내는 등 여러 가지로 활용된다.

**Text mining**

텍스트 분석과 유사한 텍스트 데이터 마이닝이라고 하는 텍스트 마이닝은 텍스트에서 고품질 정보를 도출하는 과정이다. 다른 곳에서 정보를 자동으로 추출하여 이전에 알려지지 않은 새로운 유사한 정보를 컴퓨터에서 발견하는 것이 ​​포함된다. [[3]](#endnote-3) [[4]](#endnote-4)

1. Gale, D.; Shapley, L. S. "College Admissions and the Stability of Marriage". American Mathematical Monthly. 1962 [↑](#endnote-ref-1)
2. 하임 샤피라, 『n분의 1의 함정』, 2017, 반니, 이재경 역 [↑](#endnote-ref-2)
3. Alfred V. Aho, Margaret J. Corasick, "Efficient string matching: An aid to bibliographic search", Communications of the ACM, 18, 6, 333-340, 1975. [↑](#endnote-ref-3)
4. Saima Hasib, Mahak Motwani, Amit Saxena, “Importance of Aho-Corasick String Matching Algorithm in Real World Applications”, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (3), 2013 [↑](#endnote-ref-4)