creative commons

COMMONS DEED

저작자표시-비영리-동일조건변경허락 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

동일조건변경허락. 귀하가 이 저작물을 개작, 변형 또는 가공했을 경우에는, 이 저작물과 동일한 이용허락조건하에서만 배포할 수 있습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

공학석사 학위논문

텍스트 마이닝 기법을 이용한 반도체 생산 시 발생하는 이벤트 데이터와 수율의 연관성 분석

Analysis on engineer's process event log in semiconductor processing with text mining technique

2012 년 8 월

서울대학교 대학원
산업공학과 데이터마이닝 전공
김 봉 석

텍스트 마이닝 기법을 이용한 반도체 생산 시 발생하는 이벤트 데이터와 수율의 연관성 분석

Analysis on engineer's process event log in semiconductor processing with text mining technique

지도 교수 조 성 준

이 논문을 공학석사 학위논문으로 제출함
2012 년 8 월

서울대학교 대학원
공과대학(산업공학과)
김 봉 석

김봉석의 공학석사 학위논문을 인준함
2012 년 8 월

위 원 장 _____ 홍 유 석 _____ (인)

부위원장 _____ 조 성 준 _____ (인)

위 원 _____ 박 종 헌 _____ (인)

초 록

반도체 수율(yield)은 생산하는 라인의 생산 효율성 지표이며, 회사 수익에 밀접하게 관련되어 있는 지표이다. 하나의 제품을 라인에서 생산 시, 기존 수율 대비 수율 저하 현상이 발생할 경우 문제의 원인을 빠르게 찾아 조치를 취하는 것이 반도체 회사들의 경쟁력 중에 가장 중요한 항목이라고 판단되고 있다.

지금까지 엔지니어들은 이러한 현상(수율 저하)이 발생할 경우, 아래와 같은 방법으로 문제의 원인을 찾아왔다. 첫째 라인 내에서 발생한 설비 계측 값과 수율 과의 상관관계 분석, 둘째 생산 과정 도중 계측기로부터 측정한 계측 데이터와 수율 과의 연관성 분석, 셋째 공통된 문제 설비를 도출해내는 공통성 (commonality) 분석 등을 수행하여 왔다.

하지만 위의 방법들 모두 정형화(Structured)되어 있는 데이터로부터 문제의 원인을 찾아내는 방법이었고, 엔지니어들은 전산 시스템의 도움을 통해 빠른 시간 내에 문제를 해결할 수 있었다. 그러나 위에 설명한 방법들로서 문제의 원인을 찾지 못할 경우 여러 가지 추가적 분석 작업을 수행하게 되는데 그 중 하나가 MES (Manufacturing Execution System) 내에 존재하는 생산 랏(lot) 이벤트 정보를 수작업으로 분류 및 분석하는 것이었다.

생산 랏 이벤트 정보는 라인에서 웨이퍼 생산 시 특이 사항이 발생될 경우 해당 랏의 문제점과 조치행위를 저장한 정보이다. 이 데이터는 텍스트 정보로서 시스템 상에 기록되며, 텍스트 데이터의 특징 상 비정형적(unstructured) 구조를 가지고 있다. 이러한 특징

때문에 그 동안 시스템 적으로 데이터베이스 化 할 수가 없었고 따라서 엔지니어들이 이러한 이벤트 정보를 분석하기 위해 많은 시간을 소요하였다.

그러므로 본 연구의 목적은 텍스트 마이닝의 방법론을 도입하여 반도체 생산라인 내에서 발생하는 이벤트 정보를 전처리(pre-processing)하고 텍스트 정보와 수율 과의 연관성을 분석 시도하는 것이었다. 이 연구를 통해 기존에는 텍스트 정보로 저장되어 활용되기 힘들었던 이벤트 데이터를 정량적으로 구조화 할 수 있는 가능성을 확인하였고, 향후 반도체 라인 내에서 이러한 분석 기법을 활용하여 수율 안정성에 기여할 수 있음을 확인하였다.

주요어 : 텍스트마이닝, 반도체, 수율, 라인 이벤트

학　번 : 2010-23371

목 차

제 1 장 서 론 1
제 2 장 기존 연구 5
제 3 장 제안 방법 9
3.1. 핵심 키워드의 임의 선정 9
3.2. Tf-Idf 기법을 통한 키워드 선정 10
제 4 장 실험 설정 및 결과 12
4.1. 데이터 설명 12
4.2. 데이터 전처리 20
4.3. 실험 결과 27
제 5 장 결 론 34
5.1. 결론 34
5.2. 향후 연구 방향 35
참고 문헌 37
Abstract 40

표 목차

표 [1]. 텍스트 마이닝 Technologies 7
표 [2]. 분석 대상 선정 단어 set..... 10
표 [3]. 반복 단어 발생 예시..... 10
표 [4]. 수율 데이터의 기초 통계량..... 14
표 [5]. 이벤트 데이터의 발생 형태 및 특이점 예시..... 16
표 [6]. 전처리 후 확인된 단어와 발생 빈도 예시 19
표 [7]. 데이터에서 삭제된 엔지니어 명단 21
표 [8]. 신규 등록한 단어 예시 22
표 [9]. 약어 수정 예시 22
표 [10]. Tf*idf 기법으로 선택한 단어 리스트..... 26
표 [11]. 구축된 모델의 RMSE 정리..... 28
표 [12]. 구축된 모델들의 성능 개선 여부 28
표 [13]. Stepwise Regression 에서 추출된 주요 변수..... 32
표 [14]. 주요 단어 의미의 설명 예시 33

그림 목차

그림 [1]. 반도체 공정 흐름도 1
그림 [2]. 수율 (Yield) 의 정의 2
그림 [3]. 시간에 따른, 랏 에 따른 “수율” 의 변화 3
그림 [4]. 텍스트마이닝 의 예시 5
그림 [5]. 데이터 예시 : 수율 데이터 12
그림 [6]. 이벤트 데이터 예시 13
그림 [7]. 수율 데이터의 분포 15
그림 [8]. 시간에 따른 수율 데이터의 변화 15
그림 [9]. 랏 별 발생한 이벤트 의 개수 17
그림 [10]. 랏 이벤트 개수와 수율과의 상관관계 확인 18
그림 [11]. Term frequency 분포 18
그림 [12]. 이벤트 의 Document 화 과정 20
그림 [13]. Tf, idf Ploting 결과 24
그림 [14]. Tf*idf Ploting 결과 24
그림 [15]. 단어 선택 영역 25
그림 [16]. Target 과 예측 값의 Correlation chart 29
그림 [17]. Target 과 예측 값의 Correlation chart (NN) 30

제 1 장 서 론

반도체 생산 공정은 확산, 노광, 식각, 이온주입, 화학기상증착 (Chemical vapor deposition), 금속 증착, 세정, 급속열처리 (Rapid Thermal Processing) 등 공정을 반복적으로 수행하여 이루어진다. 그림[1] 에서 볼 수 있듯이 노광 → 식각 → 세정 → 화학기상증착 등의 과정을 10~30번의 차례를 걸쳐 수행하고, 이러한 수행을 제품에 따라 20 ~ 30 층 (Layer) 을 반복하여 이루어진다. 최근 제품의 복잡도가 증가하고, 공정 미세화가 진행됨에 따라 제품을 생산하는 공정 또한 나날이 복잡해지고 있다.

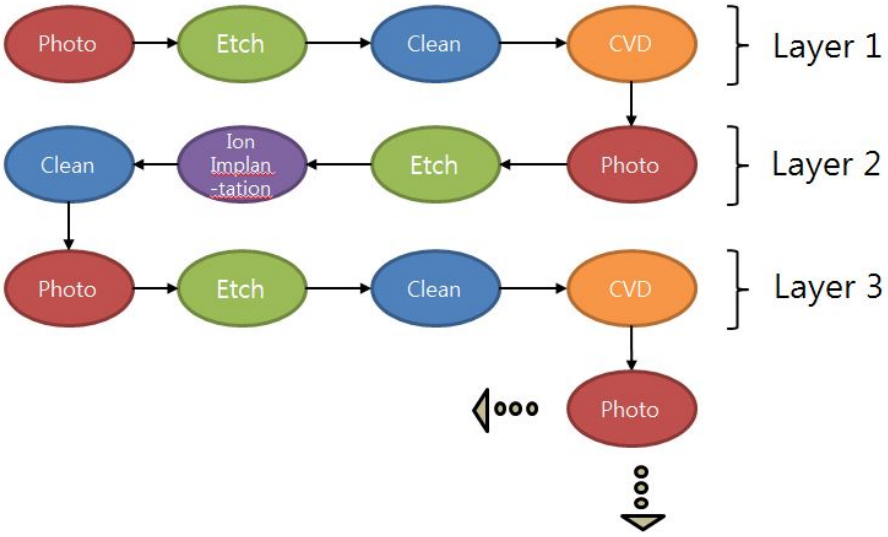
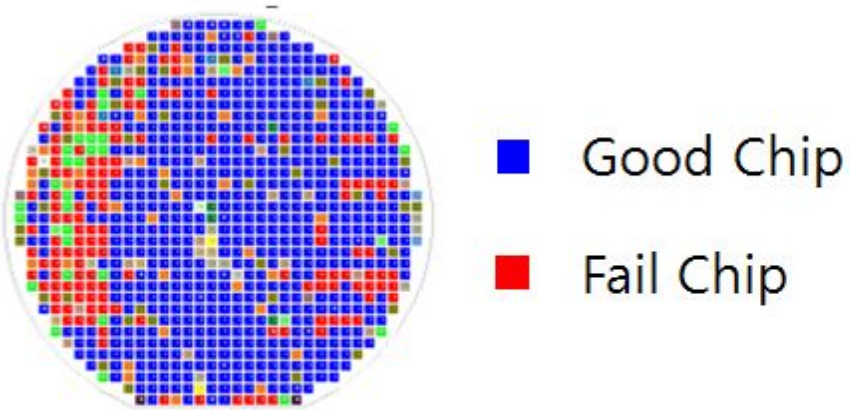


그림 [1]. 반도체 공정 흐름도

그림[2] 에서 볼 수 있듯이 반도체 생산이 끝난 후 EDS (Electronic Die Sort) 테스트를 거친 후 한 장의 웨이퍼 에서 정상

동작하는 칩 들의 비율을 수율(Yield) 라고 한다. 수율은 생산 원가 대비 얻을 수 있는 수익에 정비례하기 때문에 엔지니어들은 매 랏(웨이퍼 25매) 의 EDS 테스트 후 수율을 상시 모니터링 하고 있다.



Yield = Good Chips / Total Chips

그림 [2]. 수율 (Yield) 의 정의

그림[3] 에서 보는 바와 같이 수율은 시간의 흐름에 따라 변화하기도 하고 생산 라인 內 사고 발생에 따라 순간적으로 영향 받을 수 있다. 또한 시간의 흐름에 따른 변화는 공정 엔지니어들의 수율 개선 활동에 따른 공정 최적 조건 도출을 통해 점진적으로 상향하는 특징을 보인다. 하지만 순간적인 수율 저하나 여러 랏들의 수율 저하 현상은 반도체를 생산하는 모든 회사들이 그 원인을 찾기 위해 노력하고 있으나, 쉽게 해결할 수 없는 문제였다.

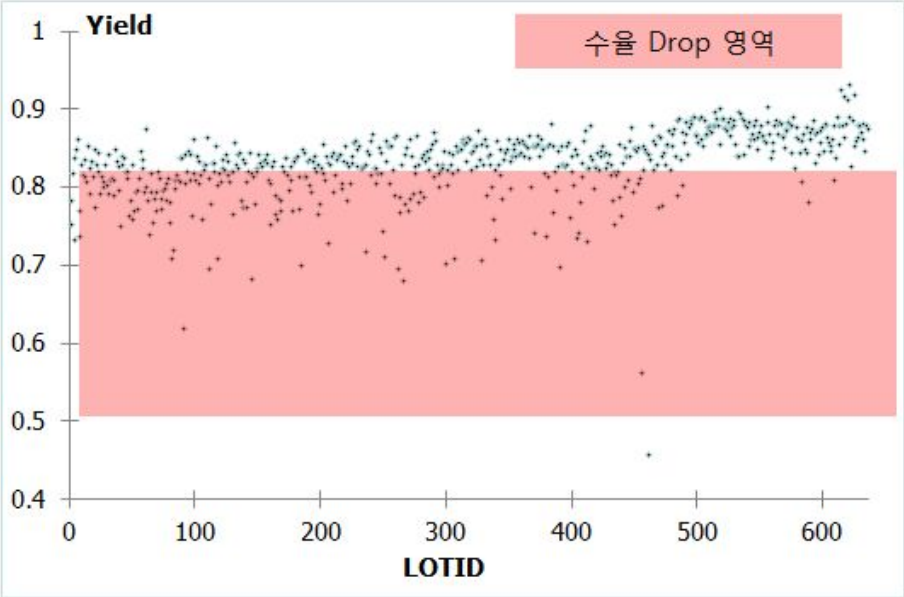


그림 [3]. 시간에 따른, 랏 에 따른 “수율” 의 변화

순간적인 수율 저하 현상의 원인을 찾기 어려웠던 이유는 아래와 같다. 첫째, 반도체 생산이 나날이 복잡해져 가기 때문에 찾아보아야 할 핵심 요소의 요소가 점점 많아진다는 점이다. 반도체 생산 시 생산 과정(Step) 사이에 각 공정의 제대로 된 생산을 확인하기 위해 계측 공정을 추가하는데 이러한 계측 정보의 데이터 양이 지속적으로 커지고 있기 때문이다. 둘째, IT 기술의 발전으로 생산에 사용되는 설비들의 실시간 데이터를 초단위로 축적하여 데이터 사이즈가 커지게 되었으나, 이를 효과적으로 분석하기 힘들기 때문이다. 셋째, 수율 저하의 원인이 계측 데이터나 설비 계측 값에 비례하는 것이 아닌 라인에 발생한 사건/사고에 유래되는 경우가 나날이 많아지고 있기 때문이다.

최근 반도체 생산이 자동화 되면서, 계측 값의 관리는 자동화되어

생산이 진행되는 동안 계측 값의 이상을 자동감지 할 수 있기에 원인 불명의 수율 저하 현상이 발생할 경우 이러한 계측 값의 추적은 유효한 결과를 얻기 힘들어졌다. 이러한 경우 엔지니어들은 여러 가지 방법으로 수율 저하의 원인을 추적 하였다. 그 중 한가지 방법은 생산 시 발생하였던 이벤트 에 대한 정보를 수작업을 통해 분류한 다음 특정 이벤트 가 수율에 영향을 미치는 지에 대해 확인하는 것이다. 하지만 이런 방법은 데이터를 처리하고 분석하는 엔지니어들의 경험에 따라 분석 성패가 나뉘는 휴리스틱(heuristic) 한 방법으로, 엔지니어의 숙련도에 따라 그 결과의 차이가 큰 단점이 있다. 게다가 이러한 수작업은 많은 시간이 소요되어 빠른 원인 탐색이 힘들었다.

하지만 위에서 설명한 수율 저하에 원인이 되는 공통 이벤트 의 확인 방법은 본 연구의 직관(Intuition)을 제공하였다. 위에서 언급한 바와 같이 이러한 비정형 이벤트 분석은 엔지니어의 숙련도에 크게 영향을 받고 있었기에, 본 연구에서는 기존의 텍스트 마이닝 기법과 데이터 마이닝 기법을 활용한 이벤트 분석 방법론을 제안하고자 한다.

제 2 장 기존 연구

텍스트 마이닝 은 기존에 알려지지 않았던 정보들을 기존에 사용하지 않았던 텍스트 정보들에서 추출하는 과정이다. 텍스트 마이닝 은 기존의 데이터 마이닝의 하나의 변형 이라 보여지고 있다. 데이터 마이닝은 정형화된 데이터 베이스 (Structured) 에서 지식을 탐색하고 패턴을 찾아내는 것을 목표로 하는 반면에 (KDD : Knowledge discovery from database) 텍스트 마이닝 혹은 Knowledge-Discovery in text (KDT) 는 비정형화된 (Unstructured data) 데이터로부터 사소하지 않고 의미가 있는 정보 (Knowledge) 를 얻어낸다는 점에서 차이가 있다 (Weiguo Fan et al. 2006). 그림[4] 에서 나타난 봐와 같이 의미 있는 정보를 얻기 위해서는 텍스트 정보에서 정형화된 데이터를 얻어야 하는데 이와 관련하여 제안된 기법들은 군집화, 시맨틱 분석 등이 있다 (Weiguo Fan et al. 2006).

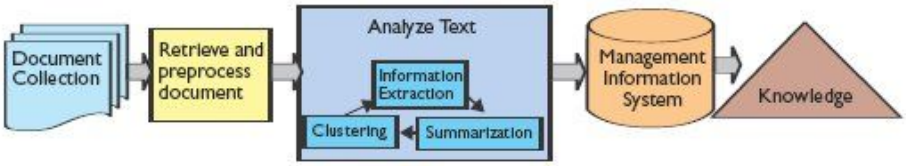


그림 [4]. 텍스트마이닝 의 예시

데이터 마이닝 은 그 동안 반도체 생산의 순익 (Profitability) 을 증대 시키기 위한 수율 증대에 큰 역할을 해왔다 . 예를 들어 Murthy (1998) 의 경우 반도체 공정의 특성상 반복되는 공정의

수율 저하와 같은 문제를 해결하기 위해서는 계층 나무 모델의 구축이 큰 효과를 가져올 수 있다고 하였으며, Bergeret and Gall(2003) 는 반도체 현장의 수율 이슈를 해결하기 위하여 베이지안 네트워크를 도입하여 공정의 사고(Failure) 가 발생하는 근본 원인이 특정한 공정에서 실패하는 것이 가장 큰 영향을 미치는 것으로 밝기기도 하였다.

현재 텍스트 마이닝 은 비정규화되지 않는 회사 내부의 정보들에서 의미 있는 정보들을 추출할 때 큰 도움이 되는 것으로 밝혀지고 있다 (Choudhary et al. 2008). 그러나 반도체 공정의 수율 안정화를 위해 도입한 데이터 마이닝 기법들 중에는 텍스트 마이닝 의 도입 연구 사례가 전무한 실정이다. 따라서 이 항목은 기존의 텍스트 마이닝 의 기초 사례를 중심으로 서술하고자 한다.

Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang (2005)은 2001년의 DOW chemical Co. 의 사례를 통해 텍스트 마이닝의 성공 사례를 예로 들었다. 약 100,000 개의 document 를 저장하고, 처리하여 10~15% 의 오차로 분류에 성공하여 신규 제품의 개발에 예전 사례 정보를 이용하는 시스템을 구축하는 사례였다. 비 구조적인 (Unstructured) 데이터의 활용이 미래에 회사들의 지식 정보를 이용한 더 나은 해결 방안이 될 수 있으리라고 예측하였고, 현재 이러한 노력들이 많이 이루어지지 않고 있음을 보였다. (현재는 정형적 데이터에 집중하는 경향이 크다고 함).

저자는 텍스트 분석은 아래와 같은 기술들이 필요하다고 하며, 이러한 기술들이 현재 인간의 자연어(natural language) 와 컴퓨터 언어를 통역(Translation) 하고 정보를 추출하는데 사용된다고

하고 있다.

표 [1]. 텍스트 마이닝 Technologies

텍스트 마이닝 Technologies
Information Extraction
Topic Tracking
Summarization
Categorization
Clustering
Information Visualization
Question Answering

Jane et al. (2010) 의 경우 항공기 조립 산업에서 거대한 항공기를 조립할 때 작업자들이 기록하는 문제 기술 보고서를 텍스트 마이닝 기법을 이용하여 온톨로지 (ontology) 분석을 시도하였다. 이 연구에서는 STAT (Semantic Trend Analysis Tool) 소프트웨어를 사용하여 120 개의 문제 기술 보고서를 분석하여 온톨로지를 산출하였고, 이를 통해 항공기 조립에 발생하는 문제들을 계층적으로 정의할 수 있었고 자주 발생하는 문제들을 정의하여 효율적으로 처리할 수 있었다.

Robert et al. (2009) 는 주식시장을 예측하기 위해 경제 뉴스의 텍스트 분석을 시도하였다. 이 연구에서 S&P 500 지수와 9,211개의 뉴스 기사, 10,259,042 개의 댓글(Comment) 정보를 활용하여 bag of words, noun phrases, named entities 를 추출한 다음 이를 주가 지수데이터와 SVM 기법을 통해 예측하였다. 이 연구는 텍스트 의 정보와 수치 데이터의 비교 분석을 시도한 점에서 의미가 있었다.

Yiming and Jan (1997)은 텍스트 분류를 통계적으로 학습하는데 필요한 변수 선택 방법들에 대해 비교 평가하였다. 이러한 변수 선택 방법이 필요한 이유는 텍스트 데이터가 가지는 광범위한 단어의 범위에서 분류 수행 시 무의미한 단어들을 삭제하고 차원 축소 (dimension reduction) 하여 효율적인 계산 성능을 얻기 위함이다. 해당 연구에서 저자들은 document frequency (DF), information gain (IG), mutual information (MI), Chi-Square test (CHI), 와 term strength (TS) 기법들의 성능 테스트를 수행하였다. 저자들의 실험 결과 성능은 IG 와 chi-Square test 방법이 성능이 가장 좋은 것으로 확인하였으나, DF 로 변수 선택을 수행하여도 그 성능이 유사하며, 계산의 복잡성도 낮기 때문에 DF 가 변수 선택에 가장 효율적이었음을 제시하였다.

제 3 장 제안 방법

3.1. 엔지니어의 핵심 키워드 선정

기존 반도체 현장의 수율 분석 엔지니어가 수율 문제가 발생할 경우이벤트 분석을 시도할 시 주로 살펴보는 단어들을 선정하여 해당 단어들이 문서에 발생하는 빈도만을 가지고 수율과 비교 하는 기법을 사용하였다. 이러한 경우 분석 대상 단어의 set 를 선정할 때 마다 분석의 결과가 바뀌는 결과를 초래할 수 있고, 실제 수율 문제가 발생한 단어가 분석에 제외될 경우 현장에서 발생하는 진성 문제에 대한 분석을 할 수 없다는 단점이 있다.

그러나 이러한 방법은 실제 현장의 선험적 지식 (Domain Knowledge) 를 반영한다는 측면에서 유효한 방법이라 판단된다. 이는 엔지니어들이 수년 동안 수율 분석을 하면서 수율과 연관 관계가 있었던 단어들을 미리 알고 있고 분석에 적용할 수 있다는 장점이 있다. 또한 특정한 단어가 발생할 경우 수율에 치명적인 영향을 받는 단어들을 미리 선정함으로써 3.2 절에서 제안하는 방법을 사용할 경우 발생할 수 있는 중요 단어가 분석에 제외되는 가능성을 배제할 수 있다는 장점이 있다.

수율 분석 엔지니어와의 인터뷰를 통해 선정된 단어 군집은 아래 표[2] 와 같으며 전체 42 개의 단어가 선정되었다.

표 [2]. 분석 대상 선정 단어 set

선정된 Word Set			
sample	변경	복원	ein
sensor	thk	alarm	cln
split	thklow	haze	진성 specout
ein	thkhigh	귀책	fail
rework	2nd	pm	interlock
수율	1st	설비 변경점	erdtrend
순간정지	긴급	설비 error	erd
정지	feedback	불발통	error
실패	공정영향	확인대기	ecn
reclean	조치	재진행	후속추가확인
정전	환자		

3.2. Tf-Idf 기법을 통한 키워드 선정

기존 텍스트 마이닝 연구에서 유의미한 단어를 추출하기 위해 자주 사용되는 tf-idf 방법을 통해 단어를 선정하는 방법을 사용하였다. 그러나 단어의 발생빈도를 계산하는 방법을 기존의 방법과 다르게 적용하였다. 그 이유는 코멘트가 자동으로 생성되는 공정 단계에서 특정 텍스트가 두 번 반복적으로 발생 텍스트 상에 발생하는 특이한 구조 때문에 특정 단어의 반복 발생하게 되는데 그 예시는 아래와 같다.

표 [3]. 반복 단어 발생 예시

랏 ID	STEPSEQ	이벤트 TIME	COMMENT
1	Y0100000	2010 02 17 5:26:41 PM	재진행! 재진행!
2	Y9403000	2010 06 16 10:40:56 PM	thk 계측진행! Thk 계측진행!

위의 표[3]. 에서 보는 것과 같이 하나의 코멘트 내부에 반복되는 단어가 발생하고 있는 문제가 있어 분석 결과에 영향을 미칠 수 있다. 따라서 위의 경우에서 랏 ID 1 번의 "재진행" 이라는 단어의 발생 빈도 값을 1 로 하고 tf*idf 값을 계산하였다. 분석에 사용된 tf,idf 공식은 아래와 같다.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

$$idf_i = log\frac{|D|}{|\{d_j : t_i \in d_j\}|}$$

$n_{i,j}$: 문서 d_j 에 단어 t_i 가 나타나는 횟수

$\sum_k n_{k,j}$: 문서 d_j 내의 모든 단어가 나타나는 횟수

|D| : *전체 문서의 숫자*

$|\{d_j : t_i \in d_j\}|$: *단어 t_i 가 등장하는 문서의 숫자*

위에서 서술한 바와 같이 중복으로 발생하는 단어들에 대한 중복 횟수를 제거하고 tf*idf 값을 계산한 다음 특정 tf*idf 값을 역치(threshold)로 지정하여 단어를 선별하였다. 이에 대한 과정은 4.2 절에서 다룰 것이다.

제 4 장 실험 설정 및 결과

4.1. 데이터 설명

가. 사용 데이터 설명

사용된 데이터는 실제 삼성전자에서 2010 년 6 월 생산된 65nm 공정의 통신 칩 제품의 625 랏 의 데이터를 활용하였다. 625 랏 의 각 랏 평균 수율과 라인에서 생산할 시 발생하였던 이벤트 정보를 수집하여 분석하였다.

LOTID	Yield
1	75.12
2	78.37
3	81.72
4	83.82
5	73.18
6	84.84
7	86.04
8	73.61
9	76.89
10	82.94
.	
.	
.	
623	84.67
624	87.89
625	87.35

그림 [5]. 데이터 예시 : 수율 데이터

그림[5] 에서 보는 것과 같이 각 랏 별로 수율이 다름을 알 수 있다. 수율 데이터는 백분율로 표시된 것으로 0 에서 100 사이의 값을 갖는다.

그림[6]처럼 이벤트 데이터는 각 생산 공정 단계별로 발생하며, 기계적으로 자동 발생된 이벤트 와 엔지니어가 수작업으로 입력한 이벤트 가 혼재되어 있다.

LOTID	STEPSEQ	EVENTTIME	DESCRIPTION
1	Y0100000	2010-02-16 21:11	SPLIT$@$FL0$@$8$@$$@$
1	Y0100000	2010-02-17 16:35	검사대기$@$$@$-$@$$@$
1	Y0100000	2010-02-17 17:26	reticle 검사 대기$@$FL0$@$8$@$$@$
1	Y0100000	2010-02-25 16:20	1.기타(제조)!@!FLOW!@!FLOW
1	Y0103600	2010-02-26 11:04	[Defect의뢰] 설비BackUpPHOTO서세현KLA Bay Flow
1	Y0103600	2010-03-01 17:24	1.기타(제조)!@!재의뢰!@!재의뢰
1	Y0103600	2010-03-01 17:48	[Defect의뢰] 설비BackUpPHOTO이호연KLA Bay Flow
1	Y0103600	2010-03-03 14:47	[Defect] lotid:N2CAX.1 stepid:Y0103600 귀책부서
1	Y0107000	2010-03-01 17:31	PFSP05호기 NOZZLE 검증관련 KLA 의뢰 -PHOTO 이호연-
1	Y0304000	2010-03-10 7:50	Hot Lot Hold 취소
1	Y1309000	2010-03-13 2:18	설비 ERR로 CANCEL함$@$TD (공통)설비ERROR
2	Y1655390	2010-04-13 17:32	[Defect의뢰] 불발통Lot추가검사DIFF손영대KLA Bay Flow
2	Y1655390	2010-04-13 18:20	[Defect의뢰] 불발통Lot추가검사DIFF손영대KLA Bay Flow
2	Y1655390	2010-04-13 19:30	[Defect] reqno:201004159115 lotid:N2CDD.1 stepid:Y1655390
2	Y1655390	2010-04-13 19:31	[불발통][183904]PRE PC 발생
2	Y1655390	2010-04-14 17:24	불량발생통보close
2	Y1655390	2010-04-20 13:32	1.기타(제조)!@!검사완료!@!검사완료
2	Y2413800	2010-04-24 10:12	SPC HOLD By CFLU04-2
2	Y2413800	2010-04-24 11:00	불발통 등록으로 OCAP 자동 Release
2	Y2413800	2010-04-24 11:00	[불발통] 귀책부서 : CVD(롤림성 특이맵으로 전수검사예정)
2	Y2413800	2010-04-24 11:54	불발통 귀책 변경으로 재 Hold 합니다.
2	Y2413800	2010-05-04 9:35	2.KLA 검사 완료(공정관리)!@!LDD단 Defect!@!FLOW
2	Y2500000	2010-05-05 12:18	SPC HOLD By NOSPCPLAN
2	Y2500000	2010-05-05 12:26	불발통 등록으로 OCAP 자동 Release
2	Y2500000	2010-05-05 12:26	[불발통]Y1655STEP에서 불발통 발의된 불량으로 재등록
2	Y2500000	2010-05-06 0:05	불량발생통보close
2	Y2815015E	2010-05-09 20:14	1.기타(제조)!@!flow!@!flow
2	Y2816120	2010-05-09 20:14	OFF-Line Trackout Hold By DFRS03
2	Y2816120	2010-05-09 20:15	1.기타(제조)!@!flow!@!flow
2	Y2818015E	2010-05-10 15:43	1.기타(제조)!@!진행완료!@!flow
2	Y2818130	2010-05-10 15:44	OFF-Line Trackout Hold By DFRS03
2	Y2818130	2010-05-10 15:44	1.기타(제조)!@!flow!@!flow
2	Y3013010	2010-05-11 7:21	PM 후 진행(미진행)$@$TE (공통)우선순위 LOT 진행
2	Y3013010	2010-05-11 7:21	1.기타(제조)!@!flow!@!flow
2	Y3104810	2010-05-12 15:01	CD측정$@$FP0$@$8$@$99023753$@$

그림 [6]. 이벤트 데이터 예시

나. 각 데이터 분석

데이터의 속성에 알맞은 방법론을 사용하기 위하여 각 데이터의 특징에 대하여 분석하였다.

ㄱ. 수율 데이터 분석

표 [4]. 수율 데이터의 기초 통계량

	수율 평균	표준편차	왜도 (Skewness)	첨예도 (kurtosis)
수율	83.065	3.888	−1.050	4.462

표[4] 에서 보는 것과 같이 수율 평균은 83% 이고, 표준편차는 3.88, 정규분포 대비 오른쪽으로 치우쳐 있으며, 정규분포 대비 중앙에 집중되는 형태를 띄고 있다. 이는 그림[7]에서 보는 것과 같다.

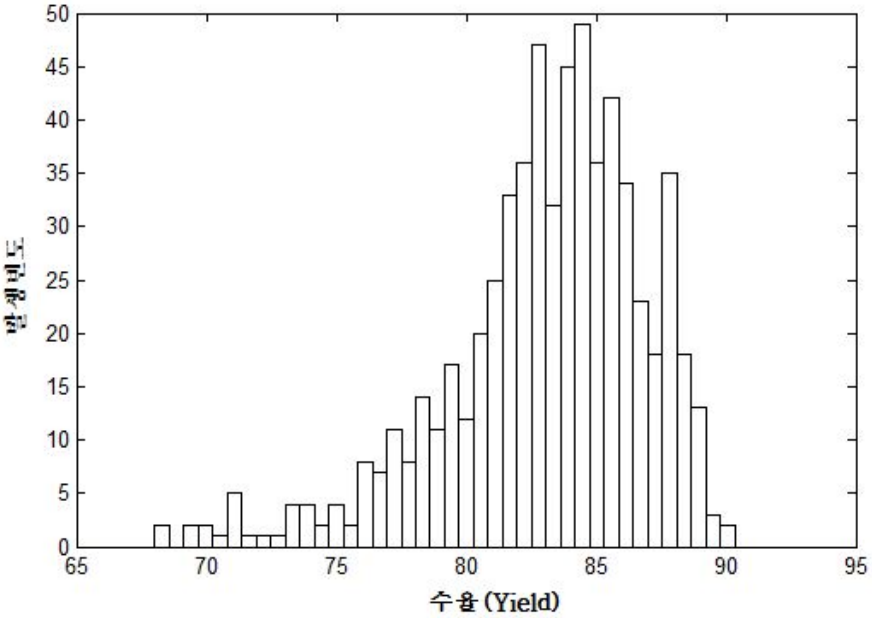


그림 [7]. 수율 데이터의 분포

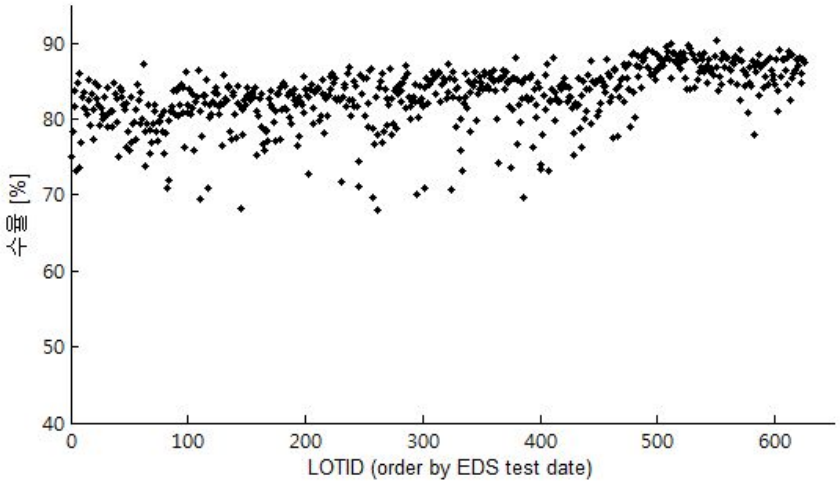


그림 [8]. 시간에 따른 수율 데이터의 변화

또한 그림[8] 에서처럼 시간에 따라 한 제품의 수율이 상승하는 경향을 보이고 있다.

ㄴ. 이벤트 데이터 분석

표 [5]. 이벤트 데이터의 발생 형태 및 특이점 예시

예시	비고
1.기타(제조)!@!FLOW!@!FLOW	전산적으로 발생한 의미없는 Comment
SPLIT$@$FL0$@$8$@$$@$	전산적으로 발생한 특수기호 모음 @,$ 등
[Defect 의뢰] 설비 BackUpPHOTO 이호연 KLA Bay Flow	띄어쓰기 없음, 띄어쓰기가 되어 있는 경우도 있음
불량발생통보 close	불발통 처럼 약어와 원래 단어가 혼용되어 쓰임
[불발통] [183904]PRE PC 발생	
R/T 검사대기$@$$@$-$@$$@$	R/T 와 Reticle 처럼 약어와 원래 단어가 혼용으로 쓰임
RETICLE 검사대기$@$$@$-$@$$@$	
$sysid$ACL ECN 위해 공법변경$@$$@$-$@07069321@$	ECN(공정개선조건 테스트) 의 진행에도 다른 코멘트가 쓰임
1.기타(제조)!@!ECN 진행!@!.	
CALCEL$@$TE (공통)우선순위 랏 진행	오타 발생 (Cancel 의 오타)
ERD	동일한 의미이나 입력하는 코멘트가 틀림
ERDERROR	
ERDTREND!@!진성 SPECOUT!@!CD 검증!@!1 중심 Trend 유지 중 1Poi	글자수 제한으로 뒤의 글자가 삭제되는 경우 발생 (Point 가 삭제됨)
RETICLE 검사대기$@$$@$-$@$$@$	동일한 의미이나 띄어쓰기가 틀린 경우
RETICLE 검사 대기$@$$@$-$@$$@$	

표[5] 에서처럼 이벤트 데이터는 사람이 직접 입력하기 때문에 오타 또는 맞춤법에 맞지 않는 경우가 많이 발생하고 있다. 아울러 전산적으로도 자동 입력이 되는 경우가 있기에 의미 없는 특수기호 (@$@# 등) 나 “1.기타(제조)!” 처럼 자동으로 입력되는 코멘트들도 많이 발생하고 있다. 이러한 데이터를 분석하기 위하여 전처리 과정을 거쳐야 함을 알 수 있었다.

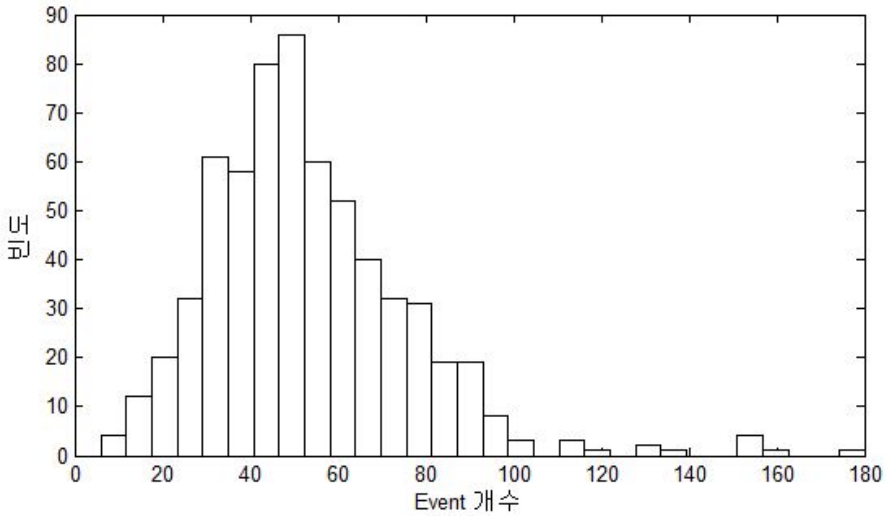


그림 [9]. 랏 별 발생한 이벤트 의 개수

그림[9] 에서 보는 것과 같이 하나의 랏 에 발생한 이벤트는 평균 50 개 내외이며, 특이한 경우 180 개 까지 이벤트가 발생하고 있다. 이벤트의 발생에 따라 수율이 발생하는지 확인하기 위해 산포도를 이용해 확인하였으나 그림[10]처럼 유의한 관계는 보이지 않았다.

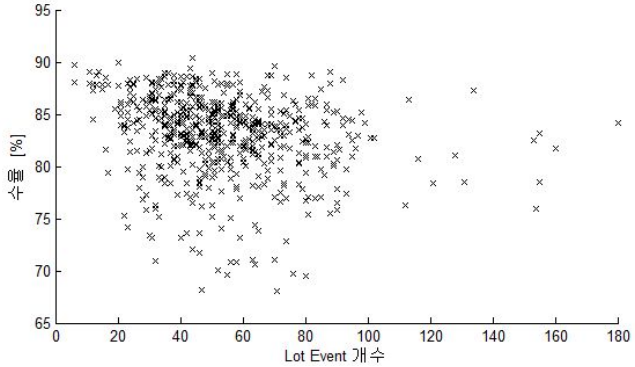


그림 [10]. 랏 이벤트 개수와 수율과의 상관관계 확인

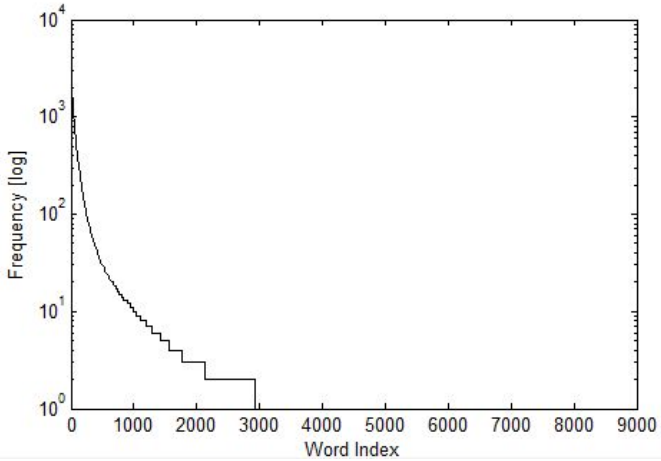


그림 [11]. Term frequency 분포

4.2 절에서 설명할 텍스트 처리 과정을 거친 후 각 단어의 발생 빈도를 확인한 결과는 그림[11] 과 같다. 특정 단어의 경우 발생빈도가 높고 한번씩 발생하는 단어의 빈도는 전체 8000 여개의 단어 세트 중에서 약 60% 이상을 차지함을 알 수 있었다. 문자 처리 과정 후에 확인된 단어와 단어 발생빈도의 예시는 아래 표[6] 와 같다.

표 [6]. 전처리 후 확인된 단어와 발생 빈도 예시

word	Word count
flow	4167
검사	3513
tracking	2971
by	2949
wip	2897
kfbk	2897
sysid	2897
귀책부서	2525
진행완료	2103
완료	2028
설비	1984
line	1882
defect	1879
hold	1874
reticle	1851

표[6] 는 추출된 단어들 중 상위 15 개의 단어의 예시를 든 것이다. 가장 많이 발생한 “Flow” 라는 단어는 라인에서 이벤트 가 발생하여 진행이 멈춘 랏 들을 다음 공정단계로 진행시키기 위해 엔지니어나 오퍼레이터들이 입력하는 단어이므로 거의 모든 코멘트에 발생하는 특징이 있다. 이러한 단어들의 필터링 기준은 4.2 절에 후술 하였다.

4.2. 데이터 전처리

수율 데이터는 숫자형 데이터이며, 결측치나 비 정상값(Outlier)이 존재하지 않았다. 따라서 별도의 전처리 과정은 거치지 않고 바로 예측 모델의 결과 변수로서 사용하였다.

이벤트 정보의 경우 발생 공정단계에 따라 하나의 랏 에 여러 개의 이벤트가 발생하고 있으나, 이를 랏 기준으로 하나로 통합하여 문서화(Document) 하였다. 이는 그림[12] 에서 보는 바와 같다.

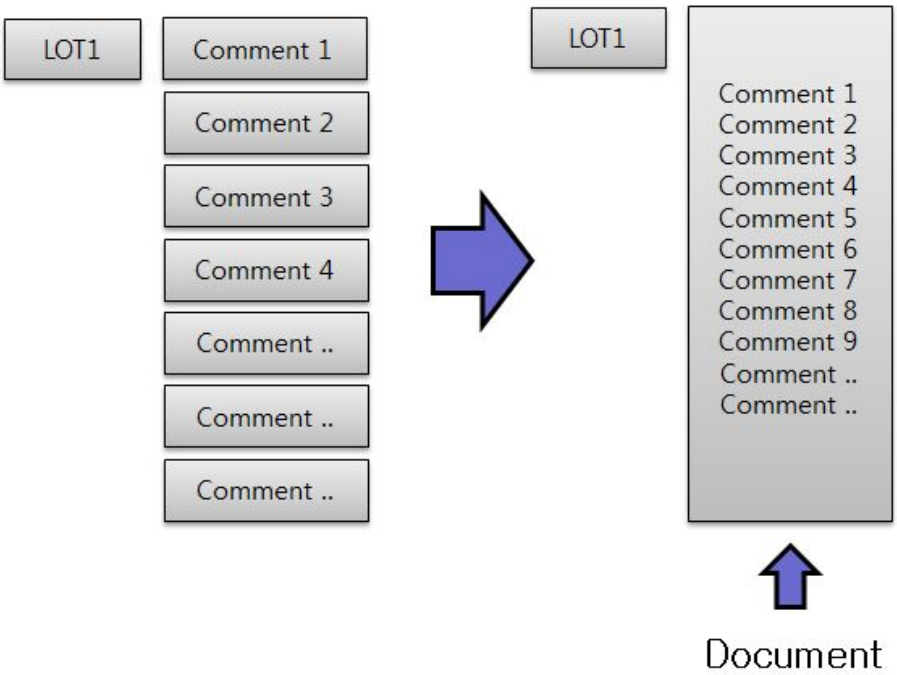


그림 [12]. 이벤트 의 Document 화 과정

코멘트들의 문서화 과정을 거친 후 아래와 같은 순서로 텍스트 정보들을 처리하였다.

(1) 데이터 입력

꼬꼬마 세종 말뭉치 Java Application 에 데이터를 로딩하였다.

(2) 데이터 필터링

첫 번째로 분석에 의미 없는 숫자형 데이터를 삭제하였다. 두 번째로 분석에 의미 없는 사람 이름을 추가로 삭제하였다. (3) 번 과정의 전문 용어 사전에 사람 (엔지니어) 이름을 추가로 입력하여 (2) 번 과정과 (3) 번 과정이 동시에 이루어지는 방식으로 진행하였다. . 삭제한 엔지니어 이름은 아래와 같다.

표 [7]. 데이터에서 삭제된 엔지니어 명단

엔지니어 명단				
변성준	송흥섭	김진규	손동진	전열
장세민	임현주	김철우	손영대	정문종
엄상용	진현종	김태성	신길용	정창윤
강일국	재현성	김태은	신용규	정호진
민은경	한준구	김해기	신원철	조광식
윤옥경	정웅택	김현철	신홍래	조양수
강위석	노태환	김효정	서세현	조오상
강은영	김태일	노원섭	안두영	조완석
강일국	김태성	류희문	안주형	지동선
강준범	주창규	박경원	양혜경	채정훈
고승억	김주광	박동환	연경진	최민희
구본송	정제령	박두호	오창건	최윤희
길현규	정주영	박민주	원종태	최인지
김태화	정미진	박범철	유영우	최현석
김다정	정찬성	박상민	은광호	한민정
김동환	전공정	박세홍	이기태	한승용
김명식	김재범	박영태	이병열	한혜선
김명철	이형우	박은미	이상헌	허혜련

세 번째로 분석 대상 데이터에 존재하는 특수 기호를 삭제하였다.

(3) 전문용어 사전 편집 및 동의어 사전 편집

반도체에서 주로 사용되는 단어들을 꼬꼬마 세종 말뭉치 프로그램에서 사용하는 사전에 추가 하였다. 추가된 단어의 예시는 아래와 같다.

표 [8]. 신규 등록한 단어 예시

신규 등록 단어	
잉여재공	관련부서
순간정지	스킵금지
순간정전	계측오류
순간변화	설비에러
공법변경	정기피엠
귀책부서	검사대기
후행진행	신규설비

두 번째로 엔지니어들이 흔히 사용하는 약어들을 하나의 형태로 통합할 수 있게 단어들을 변경하였다. 그 예시는 아래 표[9] 과 같다.

표 [9]. 약어 수정 예시

수정 전	수정 후
r/t	reticle
r.t	
reticle 검사	reticle inspection
thickness	THK
s/c	scratch
s.c	
spc hold	spchold

(4) 단편화(Tokenizing)

꼬꼬마 세종 말뭉치 프로그램을 활용하여 단어의 단편화를 실시하였다.

(5) 단어 발생 횟수 (Word count) 계산

추출된 단어들이 해당 문서에서 몇 번씩 발생하였는지 Java programming 을 통해 계산하였다.

(6) tf*idf 계산

3.2 절에서 서술한 바와 같이 분석 대상 변수를 선정하기 위해 tf*idf 값을 계산하였다. 또한 각 단어의 tf*idf 값은 분석 대상 변수인 수율을 예측하기 위한 X 인자로 사용하였다.

(7) 분석 대상 단어 선정 (feature selection)

3.1 에서 설명한 바와 같이 엔지니어들이 선험적으로 가지고 있는 중요 대상 단어를 선택하거나 3.2 에서 언급한 것처럼 tf, idf 값으로 분석 대상 단어를 선택하였다. 3.1 절과 같이 변수를 선택할 경우 별도의 데이터 전처리 과정이 필요 없었으나, 3.2 절의 방법으로 변수를 선택할 경우 역치(threshold) 를 결정해야 하는 문제가 발생한다. 따라서 적절한 변수 선택을 위해 tf, idf 데이터를 확인해 보았다.

변수 선택을 위해 그림[13]과 같이 tf,idf 을 그림[14] 와 같이 tf*idf 값을 표시하여 각 단어의 발생 빈도와 그 유효성을 확인하였다.

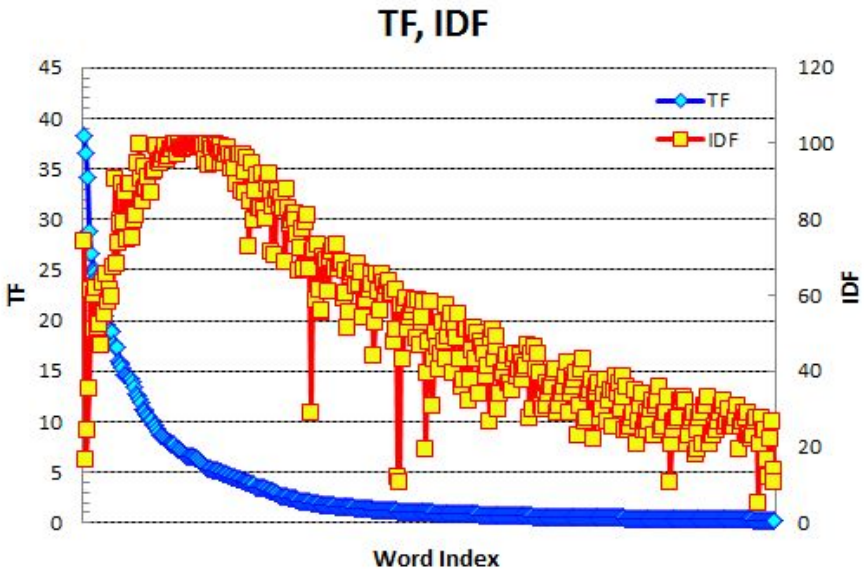


그림 [13]. Tf, idf Ploting 결과

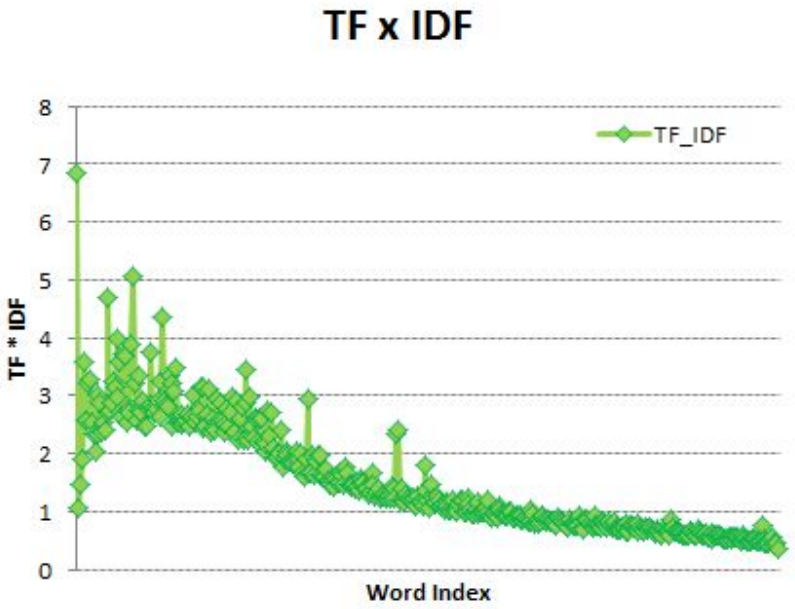


그림 [14]. tf*idf Ploting 결과

tf*idf 의 값으로 plotting 후 각 단어 별 tf*idf 값과 실제 유의미한 단어인지 수작업으로 확인을 진행한 결과 tf 가 5 이상, idf 가 20 이상, tf*idf 가 1 이상인 단어들을 고를 수 있었다. 해당 단어들을 tf*idf 차트상에 도시하면 그림[15] 와 같고 그 해당 영역은 차트에 붉은 색 박스로 표시하였다. 해당 영역에 속하는 단어의 개수는 전체 427 개 이였다.

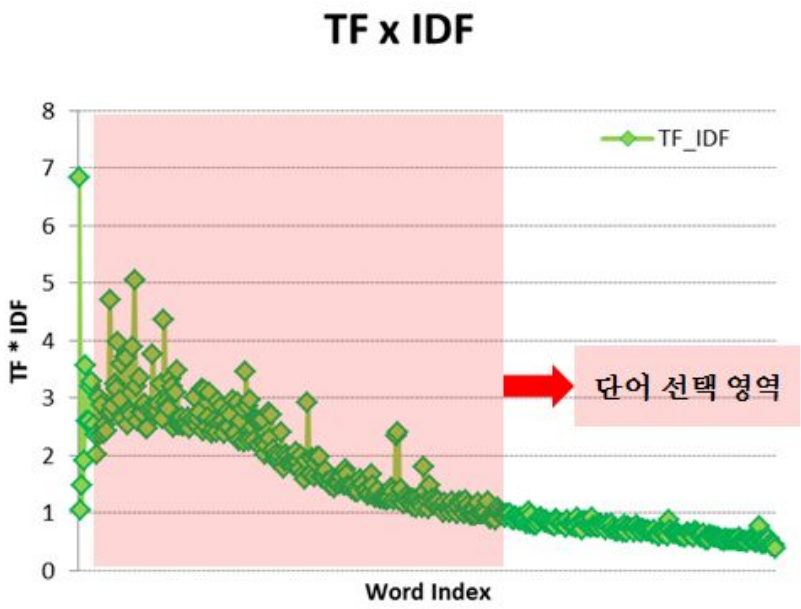


그림 [15]. 단어 선택 영역

위에서 확인한 단어들의 리스트의 예시는 표[10] 과 같으며 3.1 절에서 설명한 엔지니어가 선험적으로 알고 있는 중요단어 리스트를 전체 42 개 중 33 개를 포함함을 알 수 있었다. 실제로 엔지니어가 중요하다고 생각하는 단어들은 tf*idf 기법으로 확인할 경우 필터링 되지 않고 선택되는 것을 확인하였다.

표 [10]. Tf*idf 기법으로 선택한 단어 리스트

abnormal	cln	error	li	pcsp	secs	void	생성	정상진행
acl	close	etch	lifting	pdb	send	웨이퍼	서버	순간정전
action	cmp	exh	limit	pds	sensor	wbild	선행	정지
adh	coat	exhaust	lo	pefe	setup	wcar	설비	정체
adi	confirm	expo	local	pepc	shutdown	wcep	설비에러	제약
af	control	fab	log	pepd	shutter	wee	설정	제작사
alarm	controljob	fail	Lot clean	performance	si	wh	세정	제품
align	cot	fd	Lot cmp	pes	signal	ws	센서	조건
all	ct	fdc	Lot ids	pesp	skip	감지	수정	조정
apc	cu	feedback	lots	pfsh	skipped	개발	순간	조치
apcparameter	cvd	fem	low	pfsp	slot	검증	순간정지	중심
ar	cw	ff	lw	photo	small	결재	스킵금지	지침서
arm	cy	fh	mac	pie	sn	계측	신규	진성
arrange	데이터	fi	macro	plan	sortmgr	공법	실시	진행시
auto	dc	fix	map	planchg	spc	공정	실패	처리
autoholdrelease	deep	fj	margin	pm	spchold	공통	양호	최외각
avi	delete	fk	mcc	pn	spec	관리	에러	최적화
back	dep	flo	measurement	point	specout	교체	엔지니어	추가
backup	depo	fm	merge	polymer	sphere	귀책	영향	취소
bad	depth	fn	mes	por	split	금지	영향성	측정
barc	dev	focus	metal	pr	spot	긴급	예정	특이
bay	dg	foup	mi	pre	step	낱장	오류	특이점
bct	diff	fp	miss	pres	stepid	대기	오염	평가
beam	dme	fs	mo	present	stop	도착	完	必
before	dmedme	fu	monitoring	process	system	두께	要	필요
bf	down	future	mspc	pta	target	등록	요청	해당
bg	drop	futurehold	mx	pump	te	문제	우선순위	해제
big	dummyrunning	fv	nm	raw 데이터	temp	물량	운영	현상
bpm	dynamic	fw	normal	ray	test	미발	원인	호기
bridge	ebtc	ga	nospcplan	rcp	tf	미세	有	호환
bump	eclt	haunting	nozzle	recipe	tg	미진행	유지	확산
call	ecn	hc	np	relase	thk	발생	의뢰서	환자
cancel	ectv	high	npw	release	time	발의	이동	후속
cancle	edge	holdtime	ocap	reprocess	tkin	방향	이력	후행진행
cclk	eew	hot 랏	oecsnomsgname	reqno	tkout	변경	이상	흐름도
cclp	eexy	hs	offset	resetting	touch	변화	임의	
ccpc	efcb	ifhc	online	residue	tox	보정	자동	
cctf	efls	ifmc	open	reticle	trace	복원	자사	
cctl	eflt	inform	other	rework	track	불가	작업	
cctm	efpa	initialize	out	robot	trackin	불량	장경우	
cd	ein	interlock	over	rrc	trackout	불명	재진행	
cfas	eng	issue	overlay	rs	trend	불발통	재측정	
cflu	ep	job	ovl	run	tune	비정규	적용	
cftn	epd	kccu	pa	rx	ul	사용가능	전산	
ch	eq	kfil	pad	sample	unetch	삭제	전수	
chamber	eqp	kfsw	parameter	sbbc	uniformity	산포	전환	
check	eqpid	la	particle	sc	up	상태	점검	

4.3. 실험 결과

가. 실험 조건

텍스트 데이터의 전 처리 후, tf*idf 값을 행렬화 하였고, 수율 데이터를 병합한 다음 단계적 변수선택 회귀분석(Stepwise Linear Regression)과 단순 회귀 분석 그리고 인공신경망(Neural Network)를 통해 모델을 구축하였다. 학습(Training)에 사용되는 레코드 수와 검증(Validation) 레코드 수, 평가(Test) 에 사용하는 레코드 수의 비율은 4:3:3 으로 실험을 진행하였다. 인공신경망 모델에 사용된 은닉 노드(hidden node) 의 수는 10 개이며 학습에 사용된 알고리즘은 Levenberg-Marquardt 알고리즘이었다.

나. 실험 결과

학습된 모델의 성능을 평가하기 위하여 비교평가 지표를 설정할 필요가 있었다. 실험에 사용한 625 랏 의 수율 데이터의 기초 통계량은 표[4]에서 살펴본 것처럼 수율 평균이 83%, 표준편차가 3.88 인 값이었으므로 만약 모델을 구축하지 않고 단순히 수율 값을 83% 로 예측한다면, 그 예측의 오류를 3.88 로 예상할 수 있을 것이다. 따라서 본 연구에서는 구축된 모델의 평가지표를 수율 데이터의 표준편차인 3.88 로 두고 모델에서 예측한 값과 실제 값의 차이인 RMSE(Root Mean Squared Error)를 비교하는 것으로 모델의 성능을 평가하기로 하였다.

표[11] 에서 보는 것과 같이 구축된 모델들의 RMSE 는 3.29 에서 7.87 의 값을 갖는 것으로 확인되었다. 비교평가 지표

대비 성능이 개선되는 조건은 3.2 절에서 제안한 tf*idf 를 통한 단어 선택과 단계적 변수선택 회귀분석 기법과 인공신경망 기법 밖에 없었다. 성능의 개선 여부는 표[12]에 도시하였다.

표 [11]. 구축된 모델의 RMSE 정리

구분	Tool	Training RMSE	Test RMSE	R-Square
제안기법 1	MLR (Stepwise)	3.67	4.56	0.77
	MLR (Normal)	2.45	7.87	0.68
	Neural Network	3.45	4.48	0.81
제안기법 2	MLR (Stepwise)	2.43	3.46	0.83
	MLR (Normal)	2.22	4.08	0.67
	Neural Network	2.56	3.29	0.84

표 [12]. 구축된 모델들의 성능 개선 여부

구분	Tool	성능개선 여부	성능 개선 (%)
제안기법 1	MLR (Stepwise)	X	-14.91
	MLR (Normal)	X	-50.69
	Neural Network	X	-13.39
제안기법 2	MLR (Stepwise)	O	2.64
	MLR (Normal)	X	-2.02
	Neural Network	O	14.54

다. 실험 결과 해석

표[11] 의 결과에서 다음과 같은 사실을 유추할 수 있었다. 모델을 구축할 때 변수를 선택하지 않고 다중회귀분석을 수행할 경우 학습 시에 모델 자체가 과적합 (overfitting) 하는 경향을 보이고 있다. 이는 검증과 평가 시 RMSE 가 높아지는 현상으로 확인이 가능하다. 따라서 전 처리를 통해 분석 대상 단어들의 개수를 줄이더라도 회귀모델의 정확도를 올리기 위하여 변수 선택이 필수임을 알 수 있었다.

단계적 변수 선택 회귀분석을 사용하는 경우 제안 기법에 따라 모델의 정확도에 차이가 발생하였다. 제안 기법 1 (엔지니어가 선택한 변수만을 가지고 모델을 구축)을 적용하는 경우 제안 기법 2 에서 선택된 변수들이 선택되지 않아서 모델의 정확도가 떨어지는 것으로 유추할 수 있다.

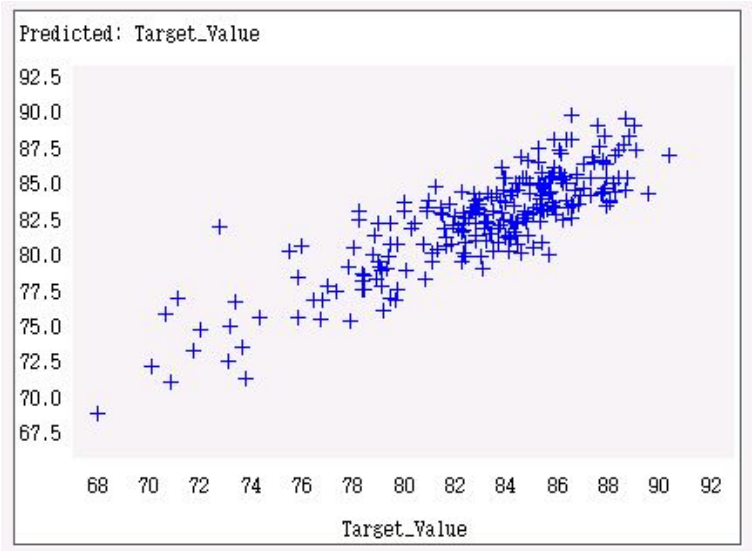


그림 [16]. Target 과 예측 값의 Correlation chart

아울러 그림[16]에서 보는 바와 같이 단계적 변수 선택 회귀분석 모델을 통한 예측은 실제 목표 값과 R-square 0.83 으로 높은 것으로 확인되어 모델이 유의미함을 알 수 있었다.

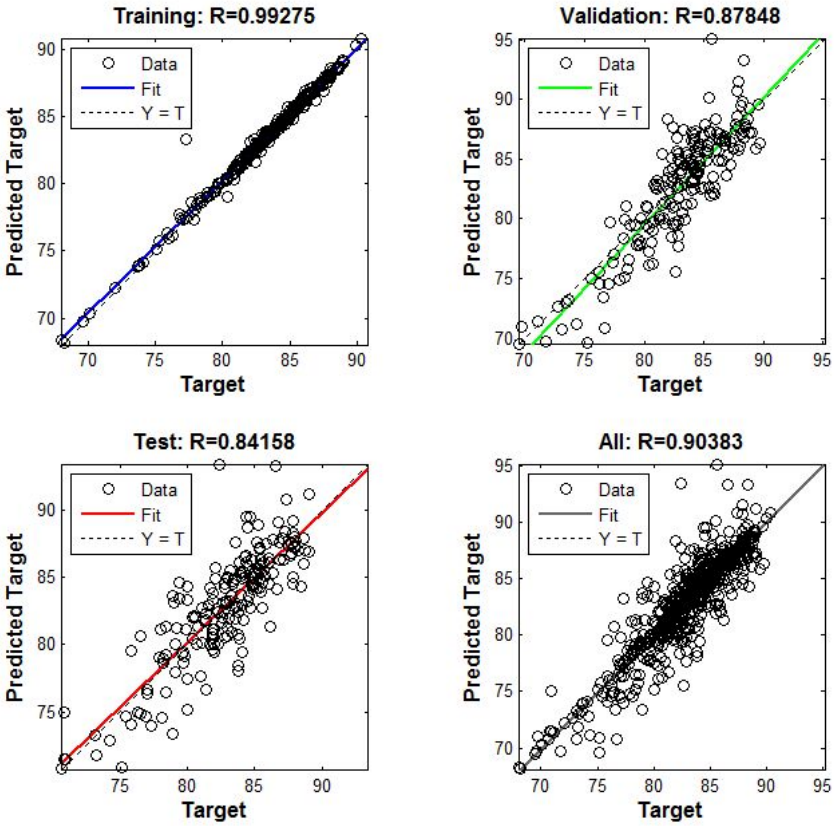


그림 [17]. Target 과 예측 값의 Correlation chart (NN)

인공신경망 모델을 통한 모델의 성능이 비교평가 지표 대비 14% 향상된 것으로 확인되었다. 인공신경망 모델은 어떠한 데이터도 적합할 수 있는 모델로서 비 선형적인 데이터 관계를 학습하는데 유리한 것으로 알려져 있다.

그러나 인공신경망 모델은 어떠한 변수가 목표 변수에 가장 크게 영향을 미치는지 알 수 없다는 단점이 있었기에, 본 연구의 목적이었던 어떤 단어가 발생할 경우 수율에 영향을 미치는지는 명확하게 알 수 없었다. 수율에 영향을 크게 미치는 텍스트 가 어떤 것이 있는지 확인해 보기 단계적 변수 선택 회귀분석의 변수 별 계수 값을 표[13]와 같이 확인해 보았다.

가장 큰 계수를 가지는 “순간변화”는 설비의 계측값이 순간적으로 변화되어 설비불량이 발생하였을 경우에 이벤트 에 남는 단어이다. 두 번째 단어인 불발통은 불량 발생 통보의 약자로서 공장 내 불량이 확인 되었을 경우 관련 부서에게 공장에 문제가 있음을 알려주는 단어이다. 위에서 설명한 두 개의 단어 모두 공장에 이상이 감지 되었을 경우 공장 작업자나 엔지니어가 이벤트 입력 시스템에 기입하는 정보이다. 그 외의 변수들 또한 공장의 이상 감지를 의미하는 단어가 모델의 영향도가 큰 것으로 알 수 있다. 이와 같은 단어들의 설명의 예시는 표[14]과 같다.

표 [13]. Stepwise Regression 에서 추출된 주요 변수

word	pval	coefficient	abs(coefficient)
순간변화	0.000	−95.560	95.560
불발통	0.001	−85.602	85.602
귀책	0.000	−85.240	85.240
inform	0.002	−82.560	82.560
call	0.000	−82.114	82.114
delete	0.004	−76.044	76.044
dmedme	0.000	−75.818	75.818
hot 랏	0.003	−63.007	63.007
설비에러	0.016	−61.563	61.563
pta	0.001	−60.247	60.247
순간정전	0.010	−59.365	59.365
cvd	0.000	−57.017	57.017
pattern	0.021	−56.173	56.173
방향	0.010	−55.192	55.192
pre	0.024	−55.123	55.123
np	0.047	−45.424	45.424
pc	0.015	−45.358	45.358
신규설비	0.002	−39.888	39.888
제품	0.029	−39.694	39.694
metal	0.020	−38.964	38.964
efpa	0.045	−38.914	38.914
edge	0.016	−36.809	36.809
send	0.000	−36.212	36.212

표 [14]. 주요 단어 의미의 설명 예시

word	의미
순간변화	설비의 parameter 가 순간적으로 바뀌어서 설비 Fault 가 발생하였을 경우
불발통	불량 발생 통보의 약자. 라인 내 불량이 확인 되었을 경우 발생하는 comment
귀책	라인내 특정 공정에서 불량이 확인될 경우 귀책부서를 기입함
inform	엔지니어에게 라인에 발생한 사고 사항을 통보하는 경우 발생
call	엔지니어에게 연락을 하라는 단어 (불량이 발생할 경우)
dmedme	Defect Management Engineer Team 의 약자로 웨이퍼 에 Particle, Defect 들이 특이한 형태로 발생할 경우 엔지니어가 기입함
Hot lot	계측을 skip 하고 생산 step 만 진행하라는 의미
설비에러	설비에 에러가 발생하여 interlock 되는 경우 발생
순간정전	순간적으로 라인에 정전이 발생하여 설비가 멈추는 경우 발생하는 comment
cvd	Chemical Vapor Deposition 의 약자로서 CVD 공정에서 에러가 발생하는 Comment 에 발생
pattern	Defect 검사 중 Pattern 이 손실된 Defect 이 발생할 경우
방향	특정 방향에 Defect 이 몰림성으로 발생할 경우 작성되는 comment 에 첨부
pre	Defect 검사 중 pre pc 가 발생할 경우
metal	Metal 공정의 약자로서 Metal 공정에서 에러가 발생하는 Comment 에 발생

제 5 장 결 론

5.1. 결론

이번 연구의 성과를 요약하면 다음과 같다.

첫째, 텍스트 마이닝 기법에서 자주 사용되는 tf*idf 를 이용한 중요 단어 선택 기법은 기존 엔지니어들이 선험적으로 알고 있던 단어의 선택과 유사함을 확인하였다.

둘째 데이터 마이닝 기법에서 사용하는 다중회귀분석, 인공신경망의 기법을 활용하여 텍스트 데이터를 기반으로 수율을 예측하는 모델을 구축할 수 있었다.

그러나 이번 연구에서 사용한 샘플 데이터의 크기가 실제 생산 현장에서 발생하는 데이터의 작은 일부에 불과하여 대량의 텍스트 데이터를 기반으로 한 현장검증을 진행할 수가 없었다.

또한 연구에서 사용한 샘플 데이터가 발생한 제품은 수율 안정화 상태에 진입하여 목표 값 자체의 분산이 매우 작은 값을 가지고 있었다. 이로 인해 텍스트 마이닝 기법을 통한 분석이 유의미한 결과를 도출하였는지 확인하기 힘들었다.

본 연구는 반도체 생산 현장에서 수율 분석을 담당하였던 선배 엔지니어들의 업무 방식에서 아이디어를 얻어 시작되었다. 숫자 데이터를 사용하여 저수율의 원인을 밝히지 못하는 경우, 라인 내 작업자들이 남기는 작업 일지를 수작업 하여 분석했던 선배 엔지니어들의 경험적 지식을 텍스트 마이닝 기법으로 구현해 보는

것이 이번 연구의 목적이었고 선배 엔지니어들이 가지고 있던 선험적 지식(domain knowledge)를 수치적으로 의미가 있음을 확인하였다는 것이 본 연구의 의의라고 할 수 있을 것이다.

5.2. 향후 연구 방향

반도체 현장에서 정형적 데이터를 사용한 수율 분석은 과거 10 년간 크게 발전해 왔다. 하지만 본 연구에서 다루었던 비정형적 텍스트 데이터를 이용한 수율 분석에 대한 연구는 전무한 실정이다. 따라서 본 연구에서 확인하였던 텍스트 데이터를 활용한 수율 분석 기법 향상에 대한 향후 연구과제의 방향은 아래와 같다.

첫째, 본 연구에서는 반도체 현장의 이벤트 데이터만 사용하여 수율과의 연관 관계를 분석하였으나 반도체 현장의 계측 수치 값과 설비 운전 데이터를 추가로 모델에 포함하여 종합적인 수율 분석 기법을 제안할 수 있다.

둘째, 라인 생산 중인 웨이퍼의 건강지수(health score)를 예측하여 수율의 저하를 미리 예측하는 모델을 구축할 수 있다. 이를 통해 EWS(Early Warning System)을 구현하여 불필요한 추가 생산 비용을 감축시킬 수 있다.

셋째, 본 연구에서 사용하였던 인공신경망과 다중회귀분석보다 예측 성능이 뛰어나다고 알려진 SVM, SVR 등을 도입하여 예측 성능을 향상 시키는 것을 시도해 볼 수 있을 것이다.

넷째, 라인에서 발생한 이벤트 데이터 외에 추가적으로 생성되는 엔지니어들의 기술 문서와 사건 사고 보고서 등을 문서 군집화나 색인을 시도해 볼 수 있다. 이러한 과제를 진행하기 위해서는 대용량의 텍스트 데이터를 처리하는 기술을 먼저 습득할 필요가 있고 대용량의 데이터베이스 또는 Hadoop 과 같은 분산 처리 시스템의 기술이 필요할 것이다.

반도체 현장 내부에서 발생하는 텍스트 정보는 공정이 발전할수록 증가하는 현상이 발생하고 있다. 본 연구과제에서 다루었던 기법들을 추가 개발할수록 반도체 수율 개선에 큰 역할을 할 수 있을 것이라 기대해본다.

참고 문헌

Bergeret, F., & Gall, C. L. (2003). Yield improvement using statistical analysis of process dates. *IEEE Transactions on Semiconductor Manufacturing, 16(3),* 535-541. doi:10.1109/TSM.2003.815204.

Berry Michael W. (2004), “Automatic Discovery of Similar Words”, in *“Survey of text mining: Clustering, Classification and Retrieval”*, *Springer Verlag, New York, LLC,* 24-43.

Choudhary, A., Harding, J., Tiwari, M., “Data mining in manufacturing: a review based on the kind of knowledge” *Journal of Intelligent Manufacturing, Springer Netherlands, Issn: 0956-5515, Business and Economics, 501-521*, vol.20, no.5, http://dx.doi.org/10.1007/s10845-008-0145-x, Doi: 10.1007/s10845-008-0145-x

Malin, J. Millward, C. Gomez, F. Throop D. (2010) , "Semantic Annotation of Aerospace Problem Reports to Support text 마이닝," *Intelligent Systems, IEEE* , vol.25, no.5, pp.20-26, Sept.-Oct. 2010 doi: 10.1109/MIS.2010.110

Murthy, Sreerama K.(1998), Automatic Construction of Decision Trees from data: A Multi-Disciplinary Survey, ***data mining and Knowledge Discovery***, Springer Netherlands, Issn: 1384-5810, Computer Science, 345-389, vol.2, no.4, Url: http://dx.doi.org/10.1023/A:1009744630224 , Doi: 10.1023/A:1009744630224

Navathe, Shamkant B., and Elmasri Ramez (2000), "Data Warehousing And data mining", in ***"Fundamentals of data base Systems"***, ***Pearson Education pvt Inc***, Singapore, 841-872.

Robert P. Schumaker and Hsinchun Chen. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFin text system. **ACM Trans. Inf. Syst. 27, 2, Article 12** , 19 pages. DOI=10.1145/1462198.1462204 http://doi.acm.org/10.1145/1462198.1462204

Yiming Yang , Jan O. Pedersen (1997), A Comparative Study on Feature Selection in text Categorization, ***Proceedings of the Fourteenth International Conference on Machine Learning***, p.412-420, July 08-12, 1997

Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang. (2006). Tapping the power of text mining. ***Commun. ACM 49, 9 (September 2006)***, 76-82. DOI=10.1145/1151030.1151032, http://doi.acm.org/10.1145/1151030.1151032

꼬꼬마 형태소 분석기 (형태소 분석기 및 자연어 처리 모듈 개발), 서울대학교 IDS(Intelligent Data Systems) 연구실, http://kkma-sc.snu.ac.kr

Abstract

Analysis on engineer's process event log in semiconductor processing with text mining Technique

Bongseok kim
Department of Industrial Engineering
The Graduate School
Seoul National University

As the indicator of efficiencies in semiconductor fabrication, semiconductor' s yield is highly correlated with firm-level profit. In semiconductor industry, it becomes accepted as a most important competitiveness that taking an immediate action to yield drop phenomenon per baseline at production line.

Commonly, most of engineers in semiconductor field are trying to find out the reason of this phenomenon, yield drop, with following methods. Correlation analysis of yield and parameters from production line's equipment, interrelation

analysis between yield and metrology data measured in measuring machine and commonality analysis to detect common error in facilities.

Because of common principle of above methods, finding out the reason of error from structured data, the engineers could find a solution to error immediately with database system. When they could not catch the reason with these methods, additional analyses need to be done. One of the further methods is manual classification and analysis of production log events from manufacturing execution system (MES).

When significant to report are detected from wafer production line, the problem of corresponding lot and action to error are saved in production log events. These events are logged in text with unstructured format. Because of characteristic of text data, the log events are could not be stored in database, so engineers should spend too much time to analysis the log events.

This paper introduces text mining method for analyzing the log events, pre-processing log events from semiconductor production line and matching interrelationship between yield and textual information from log events. From this research, I discovered the potential of quantitative structuring log events data which could have not been used to analysis. And so proposed method scheme could contribute to yield stability in semiconductor production line.

Keywords : 텍스트 마이닝, 데이터 마이닝, event log analysis,

semiconductor, semiconductor yield

Student Number : 2010-23371